A best-seller in its first edition, **Scale Development: Theory and Applications, Second Edition** has been extensively updated and revised to address changes in the field and topics that have grown in importance since the first edition. Widely adopted for graduate courses in departments such as Psychology, Public Health, Marketing, Nursing, and Education, this book will prove beneficial to applied researchers across the social sciences.

New to the Second Edition:

- Figures and practical tips for students
- New section on face validity (Chapter 4)
- Substantially expanded presentation of factor analysis (Chapter 6)
- New chapter (7) on item response theory (IRT)
- Coverage of qualitative procedures and issues related to differential item functioning (Chapter 8)

Praise for the First Edition:

"Very readable, well organized, and straightforward. I would recommend this book for practitioners, graduate students, and faculty members who are seeking a practical, rather than a psychometric, treatment of scale development. This book offers a clear overview for those interested in the development and validation of measurement scales."

—EVALUATION PRACTICE

Applied Social Research Methods Series, Volume 26
ISBN: 978-0-7619-2604-7 hardcover
ISBN: 978-0-7619-2605-4 paperback

Visit our Web site at www.sagepublications.com

ISBN 978-0-7619-2605-4
90000>
9 780761 926054

SAGE Publications
International Educational and Professional Publisher
Thousand Oaks ▪ London ▪ New Delhi

DeVellis Scale Development Second Edition

Scale Development

Theory and Applications

Second Edition

Robert F. DeVellis

Applied Social Research Methods Series
Volume 26

This book is dedicated, with deepest love and appreciation,
to my parents, John A. DeVellis and Mary DeVellis Cox,
and to my wife, Brenda DeVellis

For information:

Sage Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

Sage Publications Ltd.
6 Bonhill Street
London EC2A 4PU
United Kingdom

Sage Publications India Pvt. Ltd.
B–42 Panchsheel Enclave
New Delhi 110 017 India

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

DeVellis, Robert F.
Scale development : theory and applications / by Robert F. DeVellis.-2nd ed.
p. cm. — (Applied social research methods series ; v. 26)
Includes bibliographical references and index.
ISBN 978-0-7619-2604-7 (cloth) — ISBN 978-0-7619-2605-4 (pbk.)
1. Scaling (Social sciences) I. Title. II. Series.
H61.27 .D48 2003
300′.7′2—dc21

2002152102

09 10 11 12 13 14 11 10 9 8 7 6

Acquiring Editor: Margaret H. Seawell
Editorial Assistant: Karen Wiley
Production Editor: Claudia A. Hoffman
Copy Editor: Jamie Robinson
Indexer: Molly Hall

Contents

Preface

The first edition of this book has been widely adopted as an introduction to measurement concepts and methods. The reason for its success, I am quite sure, is that it made rather complex ideas accessible to nonexperts. That was certainly my goal. Helping students at all levels to understand measurement issues conceptually is an extremely important aspect of conveying this material. The graduate course on scale development I offer, in the School of Public Health at the University of North Carolina at Chapel Hill, attracts students with varied quantitative backgrounds. Within the same semester, my students may range from people who have had only a single graduate course in statistics to students studying for a Ph.D. in quantitative psychology. My experience in teaching that course suggests that people at all levels benefit from material presented in clear, conceptual, nonmathematical terms. Although formulas are inevitable in such a course, I try to explain the concepts in ways that make those formulas transparent, merely shorthand for a reasonable series of operations applied to data. I attempted, with some apparent success, to transfer those teaching methods to the first edition, and I have tried to do so again in this revised edition. The emphasis in this book is squarely on conveying information in ways that make the underlying principles clear and that let readers peer into the "black boxes" that various methods seem to be.

For this edition, the volume has been extensively revised. My approach has been to retain what students found clearest and most useful, to revise material for which I have conceived more lucid explanations, and to add topics that have grown in importance since the appearance of the first edition. Every chapter has changed; several chapters, substantially. More than 30 new references have been added, but many classic volumes retain their importance and are cited once again in this edition. Figures have been added to several chapters to make key points visually. In the opening chapter, I have added some new examples, clarified why some variables need many items for valid assessment and others do not, and provided a more extensive discussion of different types of item composites. Chapters 2 and 3 have been edited for greater clarity, and I have added figures illustrating key points in each chapter. In Chapter 4, in addition to doing some fine-tuning, I have added a new section on face validity. In Chapter 5, which lists the steps followed in scale development, I have added the details of several practical tips that students have found useful. Chapter 8, on viewing measurement in the broader perspective, has been expanded to include places to look for instruments, how qualitative procedures can serve as a foundation for scale

development, and issues related to differential item functioning. The two remaining chapters comprise the most substantial changes from the previous edition. Chapter 6, on factor analysis, has been substantially expanded in coverage and completely rewritten to provide more vivid and accessible analogies to the factor analytic process. I have used figures liberally to support the ideas presented textually. Finally, a new Chapter 7 has been added on a topic merely hinted at in the first edition, item response theory (IRT). My goal is not to equip readers with a working knowledge of IRT's highly complex and still-evolving methods but to provide them with a conceptual foundation that will help them understand more advanced material when they encounter it elsewhere.

Despite the inclusion of Chapter 7, the primary focus of the book continues to be classical measurement methods. There is no doubt that methods such as IRT will gain in popularity, especially as more manageable computer programs for the necessary analyses become available. Classical methods will not disappear, however. Despite certain theoretical limitations, those methods work surprisingly well in a variety of circumstances. Their foundations and applications are certainly easily understood. In various parts of the revised text, I have highlighted some areas where I think the advantages of IRT over classical methods may be particularly important. For the bulk of the research being done, however, classical methods work very well. These methods will not become obsolete as IRT gains prominence. The two will exist side-by-side as alternative methods with their respective advantages and disadvantages. Many applied researchers will never really need anything other than classical measurement techniques. The measurement gap that is still most troubling to me lies not between those who have or have not mastered the newest methods available but between those who have or have not mastered *any* measurement concepts or methodologies. I hope that this volume will help to move people to the better-informed side of that gap.

1

Overview

Measurement is of vital concern across a broad range of social research contexts. For example, consider the following hypothetical situations:

1. A health psychologist faces a common dilemma: the measurement scale she needs apparently does not exist. Her study requires that she have a measure that can differentiate between what individuals *want* to happen and what they *expect* to happen when they see a physician. Her research shows that previous studies used scales that inadvertently confounded these two ideas. No existing scales seem to make this distinction in precisely the way that she would like. Although she could fabricate a few questions to tap the distinction between what one wants and expects, she worries that "made up" items might not be reliable or valid indicators of these concepts.
2. An epidemiologist is unsure how to proceed. He is performing secondary analyses on a large data set based on a national health survey. He would like to examine the relationship between certain aspects of perceived psychological stress and health status. Although no set of items intended as a stress measure was included in the original survey, several items originally intended to measure other variables appear to tap content related to stress. It might be possible to pool these items into a reliable and valid measure of psychological stress. However, if the pooled items constituted a poor measure of stress, the investigator might reach erroneous conclusions.
3. A marketing team is frustrated in its attempts to plan a campaign for a new line of high-priced infant toys. Focus groups have suggested that parents' purchasing decisions are strongly influenced by the apparent educational relevance of toys of this sort. The team suspects that parents who have high educational and career aspirations for their infants will be most attracted to this new line of toys. Therefore, the team would like to assess these aspirations among a large and geographically dispersed sample of parents. Additional focus groups are judged to be too cumbersome for reaching a sufficiently large sample of consumers.

In each of these situations, people interested in some substantive area have come head to head with a measurement problem. None of these researchers is interested primarily in measurement per se. However, each must find a way to quantify a particular phenomenon before tackling the main research objective. In each case, "off-the-shelf" measurement tools are either inappropriate or unavailable. All the researchers recognize that, if they adopt haphazard

measurement approaches, they run the risk of yielding inaccurate data. Developing their own measurement instruments seems to be the only remaining option.

Many social science researchers have encountered similar problems. One all-too-common response to these types of problems is reliance on existing instruments of questionable suitability. Another is to assume that newly developed questionnaire items that "look right" will do an adequate measurement job. Uneasiness or unfamiliarity with methods for developing reliable and valid instruments and the inaccessibility of practical information on this topic are common excuses for weak measurement strategies. Attempts at acquiring scale development skills may lead a researcher either to arcane sources intended primarily for measurement specialists or to information too general to be useful. This volume is intended as an alternative to those choices.

GENERAL PERSPECTIVES ON MEASUREMENT

Measurement is a fundamental activity of science. We acquire knowledge about people, objects, events, and processes by observing them. Making sense of these observations frequently requires that we quantify them—that is, that we measure the things in which we have a scientific interest. The process of measurement and the broader scientific questions it serves interact with one another; the boundaries between them are often imperceptible. This happens, for example, when a new entity is detected or refined in the course of measurement or when the reasoning involved in determining how to quantify a phenomenon of interest sheds new light on the phenomenon itself. For example, Smith, Earp, and DeVellis (1995) investigated women's perceptions of battering. An a priori conceptual model based on theoretical analysis suggested that there are six distinct components to these perceptions. Empirical work aimed at developing a scale to measure these perceptions indicated that among both battered and nonbattered women, a much simpler conceptualization prevailed—a single concept thoroughly explained how study participants responded to 37 of the 40 items administered. This finding suggests that what researchers saw as a complex constellation of variables was actually perceived by women living in the community as a single, broader phenomenon. Thus, in the course of devising a means of measuring women's perceptions about battering, we discovered something new about the structure of those perceptions.

Duncan (1984) argues that the roots of measurement lie in social processes and that these processes and their measurement actually precede science: "All measurement . . . is social measurement. Physical measures are made for

social purposes" (p. 35). In reference to the earliest formal social measurement processes, such as voting, census-taking, and systems of job advancement, Duncan notes that "their origins seem to represent attempts to meet everyday human needs, not merely experiments undertaken to satisfy scientific curiosity" (p. 106). He goes on to say that similar processes "can be drawn in the history of physics: the measurement of length or distance, area, volume, weight and time was achieved by ancient peoples in the course of solving practical, social problems; and physical science was built on the foundations of those achievements" (p. 106).

Whatever the initial motives, each area of science develops its own set of measurement procedures. Physics, for example, has developed specialized methods and equipment for detecting subatomic particles. Within the behavioral and social sciences, *psychometrics* has evolved as the subspecialty concerned with measuring psychological and social phenomena. Typically, the measurement procedure used is the questionnaire, and the variables of interest are part of a broader theoretical framework.

HISTORICAL ORIGINS OF MEASUREMENT IN SOCIAL SCIENCE

Early Examples

Common sense and historical record support Duncan's claim that social necessity led to the development of measurement before science emerged. No doubt, some form of measurement has been a part of our species' repertoire since prehistoric times. The earliest humans must have evaluated objects, possessions, and opponents on the basis of characteristics such as size. Duncan (1984) cites biblical references to concerns with measurement (e.g., "A false balance is an abomination to the Lord, but a just weight is a delight," Proverbs 11:1) and notes that the writings of Aristotle refer to officials charged with checking weights and measures. Anastasi (1968) notes that the Socratic method employed in ancient Greece involved probing for understanding in a manner that might be regarded as knowledge testing. In his 1964 essay, P. H. DuBois (reprinted in Barnette, 1976) describes the use of civil service testing as early as 2200 BCE in China. Wright (1999) cites other examples of the importance ascribed in antiquity to accurate measurement, including the "weight of seven" on which 7th-century Muslim taxation was based. He also notes that some have linked the French Revolution, in part, to peasants' being fed up with unfair measurement practices.

Emergence of Statistical Methods and the Role of Mental Testing

Nunnally (1978) points out that, although systematic observations may have been going on, the absence of statistical methods hindered the development of a science of measuring human abilities until the latter half of the 19th century. Similarly, Duncan (1984) observes that, in most fields of mathematics other than geometry, applications preceded a formal development of the foundations (which he ascribes to the 19th century) by millennia. The eventual development of suitable statistical methods in the 19th century was set in motion by Darwin's work on evolution and his observation and measurement of systematic variation across species. Darwin's cousin, Sir Francis Galton, extended the systematic observation of differences to humans. A chief concern of Galton was the inheritance of anatomical and intellectual traits. Karl Pearson, regarded by many as the "founder of statistics" (e.g., Allen & Yen, 1979, p. 3), was a junior colleague of Galton's. Pearson developed the mathematical tools, including the product-moment correlation coefficient bearing his name, needed to examine systematically relationships among variables. Scientists could then quantify the extent to which measurable characteristics were interrelated. Charles Spearman continued in the tradition of his predecessors and set the stage for the subsequent development and popularization of factor analysis in the early 20th century. It is noteworthy that many of the early contributors to formal measurement (including Alfred Binet, who developed tests of mental ability in France in the early 1900s) shared an interest in intellectual abilities. Hence much of the early work in psychometrics was applied to "mental testing."

The Role of Psychophysics

Another historical root of modern psychometrics arose from psychophysics. Attempts to apply the measurement procedures of physics to the study of sensations led to a protracted debate regarding the nature of measurement. Narens and Luce (1986) have summarized the issues. They note that in the late 19th century, Hermann von Helmholtz observed that physical attributes, such as length and mass, possessed the same intrinsic mathematical structure as did positive real numbers. For example, units of length or mass could be ordered and added just like ordinary numbers. In the early 1900s, the debate continued. The Commission of the British Association for Advancement of Science regarded fundamental measurement of psychological variables to be impossible because of the problems inherent in ordering or adding sensory perceptions. S. Smith Stevens argued that strict additivity, as would apply to length or mass, was not necessary and pointed out that individuals could make fairly consistent

ratio judgments of sound intensity. For example, they could judge one sound to be twice or half as loud as another. He argued that this ratio property enabled the data from such measurements to be subjected to mathematical manipulation. Stevens is credited with classifying measurements into nominal, ordinal, interval, and ratio scales. Loudness judgments, he argued, conformed to a ratio scale (Duncan, 1984). At about the time that Stevens was presenting his arguments on the legitimacy of scaling psychophysical measures, Louis L. Thurstone was developing the mathematical foundations of factor analysis (Nunnally, 1978). Thurstone's interests spanned both psychophysics and mental abilities. According to Duncan (1984), Stevens credited Thurstone with applying psychophysical methods to the scaling of social stimuli. Thus his work represents a convergence of what had been separate historical roots.

LATER DEVELOPMENTS IN MEASUREMENT

Evolution of Basic Concepts

As influential as Stevens has been, his conceptualization of measurement is by no means the final word. He defined measurement as the "assignment of numerals to objects or events according to rules" (cited in Duncan, 1984). Duncan (1984) challenged this definition as "incomplete in the same way that 'playing the piano is striking the keys of the instrument according to some pattern' is incomplete. Measurement is not only the assignment of numerals, etc. It is also the assignment of numerals in such a way as to correspond to *different degrees of a quality* . . . or property of some object or event" (p. 126). Narens and Luce (1986) also identified limitations in Stevens's original conceptualization of measurement and illustrated a number of subsequent refinements. However, their work underscores a basic point made by Stevens: Measurement models other than the type endorsed by the Commission (of the British Association for Advancement of Science) exist, and these lead to measurement methods applicable to the nonphysical as well as physical sciences. In essence, this work on the fundamental properties of measures has established the scientific legitimacy of the types of measurement procedures used in the social sciences.

Evolution of "Mental Testing"

Although traditionally "mental testing" (or "ability testing," as it is now more commonly known) has been an active area of psychometrics, it is not

a primary focus of this volume. Many of the advances in that branch of psychometrics are less commonly and perhaps less easily applied when the goal is to measure characteristics other than abilities. These advances include item-response theory. Over time, the applicability of these methods to measurement contexts other than ability assessment has become more apparent, and we will briefly examine them in a later chapter. Primarily, however, I will emphasize the "classical" methods that largely have dominated the measurement of social and psychological phenomena other than abilities.

Broadening the Domain of Psychometrics

Duncan (1984) notes that the impact of psychometrics in the social sciences has transcended its origins in the measurement of sensations and intellectual abilities. Psychometrics has emerged as a methodological paradigm in its own right. Duncan supports this argument with three examples of the impact of psychometrics: (1) the widespread use of psychometric definitions of reliability and validity, (2) the popularity of factor analysis in social science research, and (3) the adoption of psychometric methods for developing scales measuring an array of variables far broader than those with which psychometrics was initially concerned (p. 203). The applicability of psychometric concepts and methods to the measurement of diverse psychological and social phenomena will occupy our attention for the remainder of this volume.

THE ROLE OF MEASUREMENT IN THE SOCIAL SCIENCES

The Relationship of Theory to Measurement

The phenomena we try to measure in social science research often derive from theory. Consequently, theory plays a key role in how we conceptualize our measurement problems. Of course, many areas of science measure things derived from theory. Until a subatomic particle is confirmed through measurement, it is merely a theoretical construct. However, theory in psychology and other social sciences is different from theory in the physical sciences. In the social sciences, scientists tend to rely on numerous theoretical models that concern rather narrowly circumscribed phenomena, whereas in the physical sciences, the theories scientists use are fewer in number and more comprehensive in scope. Festinger's (1954) social comparison theory, for example, focuses on a rather narrow range of human experience: the way people

evaluate their own abilities or opinions by comparing themselves to others. In contrast, physicists continue to work toward a grand unified field theory that will embrace all of the fundamental forces of nature within a single conceptual framework. Also, the social sciences are less mature than physical sciences, and their theories are evolving more rapidly. Measuring elusive, intangible phenomena derived from multiple, evolving theories poses a clear challenge to social science researchers. Therefore, it is especially important to be mindful of measurement procedures and to recognize fully their strengths and shortcomings.

The more researchers know about the phenomena in which they are interested, the abstract relationships that exist among hypothetical constructs, and the quantitative tools available to them, the better equipped they are to develop reliable, valid, and usable scales. Detailed knowledge of the specific phenomenon of interest is probably the most important of these considerations. For example, social comparison theory has many aspects that may imply different measurement strategies. One research question might require operationalizing social comparisons as relative preference for information about higher- or lower-status others, while another might dictate ratings of self relative to the "typical person" on various dimensions. Different measures capturing distinct aspects of the same general phenomenon (e.g., "social comparison") thus may not yield convergent results (DeVellis et al., 1991). In essence, the measures are assessing different variables despite the use of a common variable name in their descriptions. Consequently, developing a measure that is optimally suited to the research question requires understanding the subtleties of the theory.

Different variables call for different assessment strategies. Number of tokens taken from a container, for example, can be observed directly. Many—arguably, most—of the variables of interest to social and behavioral scientists are not directly observable; beliefs, motivational states, expectancies, needs, emotions, and social role perceptions are but a few examples. Certain variables cannot be directly observed but can be determined by research procedures other than questionnaires. For example, although cognitive researchers cannot directly observe how individuals organize information about gender into their self schemas, they may be able to use recall procedures to make inferences about how individuals structure their thoughts about self and gender. There are many instances, however, in which it is impossible or impractical to assess social science variables with any method other than a paper-and-pencil measurement scale. This is often, but not always, the case when we are interested in measuring theoretical constructs. Thus, an investigator interested in measuring androgyny may find it far easier to do so by means of a carefully developed questionnaire than by some alternative procedure.

Theoretical and Atheoretical Measures

At this point, we should acknowledge that although this book focuses on measures of theoretical constructs, not all paper-and-pencil assessments need be theoretical. Sex and age, for example, can be ascertained from self-report by means of a questionnaire. Depending on the research question, these two variables can be components of a theoretical model or simply part of a description of a study's participants. Some contexts in which people are asked to respond to a list of questions using a paper-and-pencil format, such as an assessment of hospital patient meal preferences, have no theoretical foundation. In other cases, a study may begin atheoretically but result in the formulation of theory. For example, a market researcher might ask parents to list the types of toys they have bought for their children. Subsequently the researcher might explore these listings for patterns of relationships. Based on the observed patterns of toy purchases, the researcher may develop a model of purchasing behavior. Other examples of relatively atheoretical measurement are public opinion questionnaires. Asking people which brand of soap they use or for whom they intend to vote seldom involves any attempt to tap an underlying theoretical construct. Rather, the interest is in the subject's response per se, not in some characteristic of the person it is presumed to reflect.

Distinguishing between theoretical and atheoretical measurement situations can be difficult at times. For example, seeking a voter's preference in presidential candidates as a means of predicting the outcome of an election amounts to asking a respondent to report his or her behavioral intention. An investigator may ask people how they plan to vote not out of an interest in voter decision-making processes, but merely to anticipate the eventual election results. If, on the other hand, the same question is asked in the context of examining how attitudes toward specific issues affect candidate preference, a well-elaborated theory may underlie the research. The information about voting is intended in this case not to reveal how the respondent will vote but to shed light on individual characteristics. In these two instances, the relevance or irrelevance of the measure to theory is a matter of the investigator's intent, not the procedures used. Readers interested in learning more about constructing survey questionnaires that are not primarily concerned with measuring hypothetical constructs are referred to Converse and Presser (1986), Czaja and Blair (1996), Dillman (2000), Fink (1995), Fowler (1993, 1995), and Weisberg, Krosnick, and Bowen (1996).

Measurement Scales

Measurement instruments that are collections of items combined into a composite score, and intended to reveal levels of theoretical variables not

readily observable by direct means, are often referred to as *scales*. We develop scales when we want to measure phenomena that we believe to exist because of our theoretical understanding of the world, but that we cannot assess directly. For example, we may invoke depression or anxiety as an explanation for behaviors we observe. Most theoreticians would agree that depression or anxiety is not equivalent to the behavior we see, but underlies it. Our theories suggest that these phenomena exist and that they influence behavior, but that they are intangible. Sometimes, it may be appropriate to infer their existence from their behavioral consequences. However, at other times, we may not have access to behavioral information (such as when we are restricted to mail survey methodologies), may not be sure how to interpret available samples of behavior (such as when a person remains passive in the face of an event that most others would react to strongly), or may be unwilling to assume that behavior is isomorphic with the underlying construct of interest (such as when we suspect that crying is the result of joy rather than sadness). In instances in which we cannot rely on behavior as an indication of a phenomenon, it may be useful to assess the construct by means of a carefully constructed and validated scale.

Even among theoretically derived variables, there is an implicit continuum ranging from relatively concrete and accessible phenomena to relatively abstract and inaccessible phenomena. Not all phenomena will require multi-item scales. Age and gender certainly have relevance to many theories but rarely require a multi-item scale for accurate assessment. People know their age and gender. These variables, for the most part, are linked to concrete, relatively unambiguous characteristics (e.g., morphology) or events (e.g., date of birth). Unless some special circumstance such as a neurological impairment is present, respondents can retrieve information about their age and gender from memory quite easily. They can respond with a high degree of accuracy to a single question assessing a variable such as these. Ethnicity arguably is a more complex and abstract variable than is age or gender. It typically involves a combination of physical, cultural, and historical factors. As a result, it is less tangible—more of a social construction—than is age or gender. Although the mechanisms involved in defining one's ethnicity may be complex and unfold over an extended period of time, most individuals have arrived at a personal definition and can report their ethnicity with little reflection or introspection. Thus, a single variable may suffice for assessing ethnicity under most circumstances. Many other theoretical variables, however, require a respondent to reconstruct, interpret, judge, compare, or evaluate less accessible information. For example, measuring how married people believe their lives would be different if they had chosen a different spouse probably would require substantial mental effort, and one item may not capture the complexity of the

phenomenon of interest. Under conditions such as these, a scale may be the appropriate assessment tool. Multiple items may capture the essence of such a variable with a degree of precision that a single item could not attain. It is precisely this type of variable—one that is not directly observable and that involves thought on the part of the respondent—that is most appropriately assessed by means of a scale.

A scale should be contrasted with other types of multi-item measures that yield a composite score. The distinctions among these different types of item composites is of both theoretical and practical importance, as later chapters of this book will reveal. As the terms are used in this volume, a *scale* consists of what Bollen (1989, pp. 64-65; see also Loehlin, 1998, pp. 200-202) terms "effect indicators"—that is, items whose values are caused by an underlying construct (or "latent variable," as I shall refer to it in the next chapter). A measure of depression often conforms to the characteristics of a scale, with the responses to individual items sharing a common cause, namely, the affective state of the respondent. Thus, how someone responds to items such as "I feel sad" and "My life is joyless" probably is largely determined by that person's feelings at the time. I will use the term *index*, on the other hand, to describe sets of items that are "cause indicators," that is, items that determine the level of a construct. A measure of presidential candidate appeal, for example, might fit the characteristics of an index. The items might assess a candidate's geographical residence, family size, physical attractiveness, ability to inspire campaign workers, and potential financial resources. Although these characteristics probably do not share any common cause, they might all share an effect—increasing the likelihood of a successful presidential campaign. The items are not the result of any one thing, but they determine the same outcome. A more general term for a collection of items that one might aggregate into a composite score is *emergent variable* (e.g., Cohen, Cohen, Teresi, Marchi, & Velez, 1990), which includes collections of entities that share certain characteristics and can be grouped under a common category heading. Grouping characteristics together, however, does not necessarily imply any causal linkage. Sentences beginning with a word that has fewer than five letters, for example, can easily be categorized together although they share neither a common cause nor a common effect. An emergent variable "pops up" merely because someone or something (such as a data analytic program) perceives some type of similarity among the items in question.

All Scales Are Not Created Equal

Regrettably, not all item composites are developed carefully. For many, *assembly* may be a more appropriate term than *development*. Researchers often "throw together" or "dredge up" items and assume they constitute a suitable scale. These researchers may give no thought to whether the items share a common cause (thus constituting a scale), share a common consequence (thus constituting an index), or merely are examples of a shared superordinate category that does not imply either a common causal antecedent or consequence (thus constituting an emergent variable).

A researcher not only may fail to exploit theory in developing a scale but also may reach erroneous conclusions about a theory by misinterpreting what a scale measures. An unfortunate but distressingly common occurrence is the researcher coming to the conclusion that some *construct* is unimportant or that some *theory* is inconsistent, based on the performance of a *measure* that may not reflect the variable assumed by the investigator. Why might this happen? Rarely in research do we examine relationships among variables directly. As noted earlier, many interesting variables are not directly observable, a fact we can easily forget. More often, we assess relationships among proxies (such as scales) that are intended to represent the variables of interest. The observable proxy and the unobservable variable may become confused. For example, variables such as blood pressure and body temperature, at first consideration, appear to be directly observable; but what we actually observe are proxies such as a column of mercury. Our conclusions about the variables assume that the observable proxies are very closely linked to the underlying variables they are intended to represent. Such is the case for a thermometer; we describe the level of the mercury in a thermometer as "the temperature," even though, strictly speaking, it is merely a visible manifestation of temperature (i.e., thermal energy). In this case, where the two correspond very closely, the consequences of referring to the measurement (the scale value that the mercury attains) as the variable (the amount of thermal energy) are nearly always inconsequential. When the relationship between the variable and its indicator is weaker than in the thermometer example, confusing the measure with the phenomenon it is intended to reveal can lead to erroneous conclusions. Consider a hypothetical situation in which an investigator wishes to perform a secondary analysis on an existing data set. Let us assume that our investigator is interested in the role of social support on subsequent professional attainment. The investigator observes that the available data set contains a wealth of information on subjects' professional status over an extended period of time and that subjects were asked whether they were married. In fact, there may be several items, collected at various times, that pertain to marriage. Let us further assume that, in the absence of any data providing a more detailed assessment of social support, the investigator decides to sum these marriage items into a "scale" and to use this as a measure of support. Most social scientists would agree that equating social support with marital status is not

justified. The latter both omits important aspects of social support (e.g., the perceived quality of support received) and includes potentially irrelevant factors (e.g., status as an adult versus a child at the time of measurement). If this hypothetical investigator concluded, on the basis of having used this assessment method, that social support played no role in professional attainment, that conclusion might be completely wrong. In fact, the comparison was between marital status and professional attainment. Only if marital status actually indicated level of support would the conclusion be valid.

Costs of Poor Measurement

Even if a poor measure is the only one available, the costs of using it may be greater than any benefits attained. It is rare in the social sciences for there to be situations in which an immediate decision must be made in order to avoid dire consequences and one has no other choice but to make do with the best instruments available. Even in these rare instances, however, the inherent problems of using poor measures to assess constructs do not vanish. Using a measure that does not assess what one presumes it assesses can lead to wrong decisions. Does this mean that we should only use measurement tools that have undergone rigorous development and extensive validation testing? Not necessarily. Although imperfect measurement may be better than no measurement at all in some situations, we should *recognize* when our measurement procedures are flawed and temper our conclusions accordingly.

Often, an investigator will consider measurement secondary to the important scientific issues that motivate a study and thus attempt to "economize" by skimping on measurement. However, adequate measures are a necessary condition for valid research. Investigators should strive for an isomorphism between the theoretical constructs in which they have an interest and the methods of measurement they use to operationalize them. Poor measurement imposes an absolute limit on the validity of the conclusions one can reach. For an investigator who prefers to pay as little attention as possible to measurement—and as much attention as possible to substantive issues—an appropriate strategy might be to get the measurement part of the investigation correct from the very beginning so that it can be taken more or less for granted thereafter.

A researcher also can falsely economize by using scales that are too brief in the hope of reducing the burden on respondents. Choosing a questionnaire that is too brief to be reliable is a bad idea no matter how much respondents appreciate its brevity. A reliable questionnaire that is completed by half of the respondents yields more information than an unreliable questionnaire that is

completed by all respondents. If you cannot determine what the data mean, the amount of information collected is irrelevant. Consequently, respondents' completing "convenient" questionnaires that cannot yield meaningful information is a poorer use of their time and effort than their completing a somewhat longer version that produces valid data. Thus, using inadequately brief assessment methods may have ethical as well as scientific implications.

SUMMARY AND PREVIEW

This chapter has stressed that measurement is a fundamental activity in all branches of science, including the behavioral and social sciences. *Psychometrics*, the specialty area of the social sciences that is concerned with measuring social and psychological phenomena, has historical antecedents extending back to ancient times. In the social sciences, theory plays a vital role in the development of measurement *scales*, which are collections of items that reveal the level of an underlying theoretical variable. However, not all collections of items constitute scales in this sense. Developing scales may be more demanding than selecting items casually; however, the costs of using "informal" measures usually greatly outweigh the benefits.

The following chapters cover the rationale and methods of scale development in greater detail. Chapter 2 explores the "latent variable," the underlying construct that a scale attempts to quantify, and it presents the theoretical bases for the methods described in later chapters. Chapter 3 provides a conceptual foundation for understanding reliability and the logic underlying the reliability coefficient. The fourth chapter reviews validity, while the fifth is a practical guide to the steps involved in scale development. Chapter 6 introduces factor analytic concepts and describes their use in scale development. Chapter 7 is a conceptual overview of an alternative approach to scale development, item response theory. Finally, Chapter 8 briefly discusses how scales fit into the broader research process.

2

Understanding the Latent Variable

This chapter presents a conceptual schema for understanding the relationship between measures and the constructs they represent, though it is not the only framework available. Item response theory (IRT) is an alternative measurement perspective that we will examine in Chapter 7. Because of its relative conceptual and computational accessibility and wide usage, I emphasize the classical measurement model, which assumes that individual items are comparable indicators of the underlying construct.

CONSTRUCTS VERSUS MEASURES

Typically, researchers are interested in constructs rather than items or scales per se. For example, a market researcher measuring parents' aspirations for their children would be more interested in intangible parental sentiments and hopes about what their children will accomplish than in where those parents place marks on a questionnaire. However, recording responses to a questionnaire may, in many cases, be the best method of assessing those sentiments and hopes. Scale items are usually a means to the end of construct assessment. In other words, they are necessary because many constructs cannot be assessed directly. In a sense, measures are proxies for variables that we cannot directly observe. By assessing the relationships between measures, we infer, indirectly, the relationships between constructs. In Figure 2.1, for example, although our primary interest is the relationship between variables A and B, we estimate it on the basis of the relationship between measures corresponding to those variables.

The underlying phenomenon or construct that a scale is intended to reflect is often called the *latent variable.* Exactly what is a latent variable? Its name reveals two chief features. Consider the example of parents' aspirations for children's achievement. First, it is *latent* rather than manifest. Parents' aspirations for their children's achievement are not directly observable. In addition, the construct is *variable* rather than constant—that is, some aspect of it, such as its strength or magnitude, changes. Parents' aspirations for their children's achievement may vary with regard to time (e.g., during the child's infancy versus adolescence), place (e.g., on an athletic field versus a classroom),

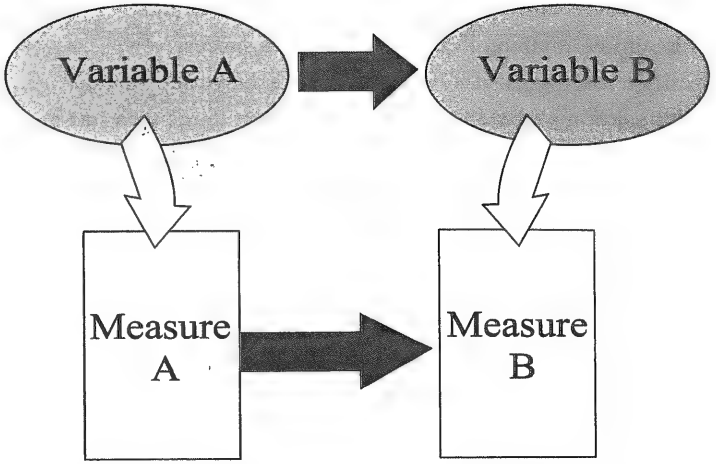

Figure 2.1 Relationships between instruments correspond to relationships between latent variables only when each measure corresponds to its latent variable

people (e.g., parents whose own backgrounds or careers differ), or any combination of these and other dimensions. The latent variable is the actual phenomenon that is of interest, in this case, child achievement aspirations. Although we cannot observe or quantify it directly, the latent variable presumably takes on a specific value under some specified set of conditions. A scale developed to measure a latent variable is intended to estimate its actual magnitude at the time and place of measurement for each person measured. This unobservable "actual magnitude" is the *true score.*

LATENT VARIABLE AS THE PRESUMED CAUSE OF ITEM VALUES

The notion of a latent variable implies a certain relationship between it and the items that tap it. The latent variable is regarded as a *cause* of the item score—that is, the strength or quantity of the latent variable (i.e., the value of its true score) is presumed to cause an item (or set of items) to take on a certain value.

An example may reinforce this point: The following are hypothetical items for assessing parents' aspirations for children's achievement:

1. My child's achievements determine my own success.
2. I will do almost anything to ensure my child's success.
3. No sacrifice is too great if it helps my child achieve success.
4. My child's accomplishments are more important to me than just about anything else I can think of.

If parents are given an opportunity to express how strongly they agree with each of these items, their underlying aspirations for their children's achievement should influence their responses. In other words, each item should give an indication of the strength of the latent variable, aspirations for children's achievement. The score obtained on the item is caused by the strength or quantity of the latent variable for that person, at that particular time.

A causal relationship between a latent variable and a measure implies certain empirical relationships. For example, if an item value is caused by a latent variable, then there should be a correlation between that value and the true score of the latent variable. Because we cannot directly assess the true score, we cannot compute a correlation between it and the item. However, when we examine a set of items that are presumably caused by the same latent variable, we can examine their relationships to one another. So, if we had several items like the ones above measuring parental aspirations for child achievement, we could look directly at how they correlated with one another, invoke the latent variable as the basis for the correlations among items, and use that information to infer how highly each item was correlated with the latent variable. Shortly, I will explain how all this can be learned from correlations among items. First, however, I will introduce some diagrammatic procedures to help make this explanation more clear.

PATH DIAGRAMS

Coverage of this topic here will be limited to issues pertinent to scale development. For more in-depth treatment of the topic, consult Asher (1983) or Loehlin (1998).

Diagrammatic Conventions

Path diagrams are a method for depicting *causal* relationships among variables. Although they can be used in conjunction with *path analysis*, which is a data analytic method, path diagrams have more general utility as a means of

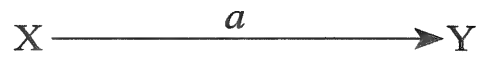

Figure 2.2 The causal path *a* from X to Y

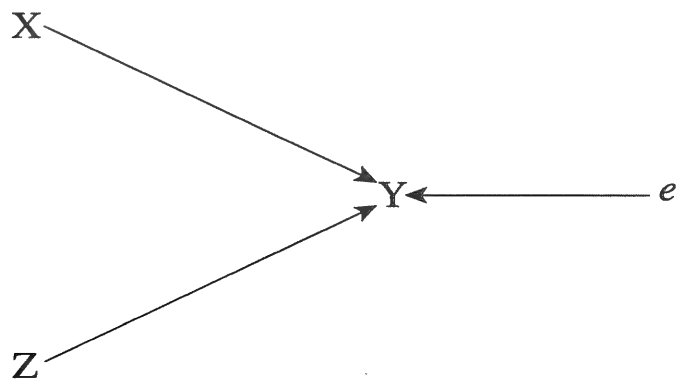

Figure 2.3 Two variables plus error determine Y

specifying how a set of variables is interrelated. These diagrams adhere to certain conventions. A *straight arrow* drawn from one variable label to another indicates that the two are *causally related* and that the direction of causality is as indicated by the arrow. Thus X → Y indicates explicitly that X is the cause of Y. Often, associational paths are identified by labels, such as the letter "a" in Figure 2.2.

The *absence* of an arrow also has an explicit meaning, namely, that two variables are *unrelated.* Thus; A → B → C D → E specifies that A causes B, B causes C, C and D are *unrelated,* and D causes E. Another convention of path diagrams is the method of representing *error,* which is usually depicted as an additional causal variable. This error term is a *residual,* representing all sources of variation not accounted for by other causes explicitly depicted in the diagram.

Because this error term is a residual, it represents the discrepancy between the actual value of Y and what we would predict Y to be, based on our knowledge of X and Z (in this case). Sometimes, the error term is assumed and thus not included in the diagram (see Figure 2.3).

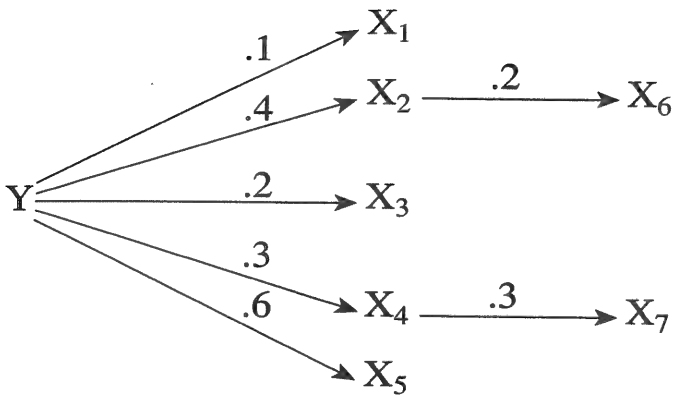

Figure 2.4 A path diagram with path coefficients, which can be used to compute correlations between variables

Path Diagrams in Scale Development

Path diagrams can help us see how scale items are causally related to a latent variable. They can also help us understand how certain relationships among items imply certain relationships between items and the latent variable. We begin by examining a simple computational rule for path diagrams. Let us look at the simple path diagram in Figure 2.4.

The numbers along the paths are *standardized path coefficients*. Each one expresses the strength of the causal relationship between the variables joined by the arrow. The fact that the coefficients are standardized means that they all use the same scale to quantify the causal relationships. In this diagram, Y is a cause of X_1 through X_5. A useful relationship exists between the values of path coefficients and the correlations between the Xs (which would represent items, in the case of a scale development-type path diagram). For diagrams like this one that have only one common origin (Y, in this case), the correlation between any two Xs is equal to the product of the coefficients for the arrows forming a route, through Y, between the X variables in question. For example, the correlation between X_1 and X_5 is calculated by multiplying the two standardized path coefficients that join them via Y. Thus, $r_{1,5} = .6 \times .1 = .06$. Variables X_6 and X_7 also share Y as a common source, but the route connecting them is longer. However, the rule still applies. Beginning at X_7, we can trace back to Y and then forward again to X_6. (Or, we could have gone in the other direction, from X_6 to X_7.) The result is: $.3 \times .3 \times .4 \times .2 = .0072$. Thus, $r_{6,7} = .0072$.

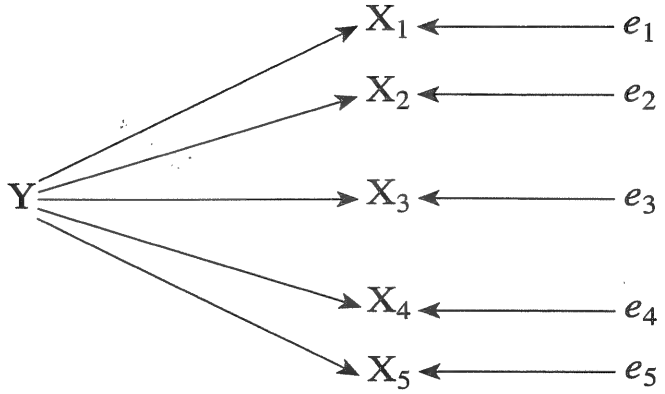

Figure 2.5 A path diagram with error terms

This relationship between path coefficients and correlations provides a basis for estimating paths between a latent variable and the items that it influences. Even though the latent variable is hypothetical and unmeasurable, the items are real and the correlations among them can be directly computed. By using these correlations, the simple rule just discussed, and some assumptions about the relationships among items and the true score, we can come up with estimates for the paths between the items and the latent variable. We can begin with a set of correlations among variables. Then, working backward from the relationship among paths and correlations, we can determine what the values of certain paths must be if the assumptions are correct. Let us consider the example in Figure 2.5.

This diagram is similar to the example considered earlier except that there are no path values, the variables X_6 and X_7 have been dropped, the remaining X variables represent scale items, and each item has a variable (error, labeled "*e*") other than Y influencing it. These *e* variables are unique in the case of each item and represent the residual variation in each item not explained by Y. This diagram indicates that all of the items are influenced by Y. In addition, each is influenced by a unique set of variables other than Y that is collectively treated as error.

This revised diagram represents how five individual items are related to a single latent variable, Y. The numerical subscripts given to the *e*s and Xs indicate that the five items are different and that the five sources of error, one for each item, are also different. The diagram has no arrows going directly from one X to another X or going from an *e* to another *e* or from an *e* to an X other

than the one with which it is associated. These aspects of the diagram represent assumptions that will be discussed later.

If we had five actual items that a group of people had completed, we would have item scores that we could then correlate with one another. The rule examined earlier allows the computations of correlations from path coefficients. With the addition of some assumptions, it also lets us compute path coefficients from correlations—that is, correlations computed from actual items can be used to determine how each item relates to the latent variable. If, for example, X_1 and X_4 have a correlation of .49, then we know that the product of the values for the path leading from Y to X_1 and the path leading from Y to X_4 is equal to .49. We know this because our rule established that the correlation of two variables equals the product of the path coefficients along the route that joins them. If we also assume that the *two path values are equal,* then they both must be .70.[1]

FURTHER ELABORATION OF THE MEASUREMENT MODEL

Classical Measurement Assumptions

The classical measurement model starts with common assumptions about items and their relationships to the latent variable and sources of error:

1. The amount of error associated with individual items varies randomly. The error associated with individual items has a mean of zero when it is aggregated across a large number of people. Thus, items' means tend to be unaffected by error when a large number of respondents complete the items.
2. One item's error term is *not* correlated with another item's error term; the only routes linking items pass through the latent variable, never through any error term.
3. Error terms are *not* correlated with the true score of the latent variable. Note that the paths emanating from the latent variable do not extend outward to the error terms. The arrow between an item and its error term aims the other way.

The first two assumptions above are common statistical assumptions that underlie many analytic procedures. The third amounts to defining "error" as the residual remaining after one considers all of the relationships between a set of predictors and an outcome, or, in this case, a set of items and its latent variable.

PARALLEL TESTS

Classical measurement theory, in its most orthodox form, is based on the assumption of parallel "tests." The term *parallel tests* stems from the fact that one can view each individual item as a "test" for the value of the latent variable. For our purposes, referring to *parallel items* would be more accurate. However, I will defer to convention and use the traditional name.

A virtue of the parallel tests model is that its assumptions make it quite easy to reach useful conclusions about how individual items relate to the latent variable, based on our observations of how the items relate to one another. Earlier, I suggested that with a knowledge of the correlations among items and with certain assumptions, one could make inferences about the paths leading from a causal variable to an item. As will be shown in the next chapter, being able to assign a numerical value to the relationships between the latent variable and the items themselves is quite important. Thus, in this section, I will examine in some detail how the assumptions of parallel tests lead to certain conclusions that make this possible.

The rationale underlying the model of parallel tests is that each item of a scale is precisely as good a measure of the latent variable as any other of the scale items. The individual items are thus *strictly parallel,* which is to say that each item's relationship to the latent variable is presumed identical to every other item's relationship to that variable *and* the amount of error present in each item is also presumed to be identical. Diagrammatically, this model can be represented as shown in Figure 2.6.

This model adds two assumptions to those listed earlier:

1. The amount of influence from the latent variable to each item is assumed to be the same for all items.
2. Each item is assumed to have the same amount of error as any other item, meaning that the influence of factors *other* than the latent variable is equal for all items.

These added assumptions mean that the correlation of each item with the true score is identical. Being able to assert that these correlations are *equal* is important because it leads to a means of determining the *value* for each of these identical correlations. This, in turn, leads to a means of quantifying reliability, which will be discussed in the next chapter.

Asserting that correlations between the true score and each item are equal requires *both* of the preceding assumptions. A squared correlation is the

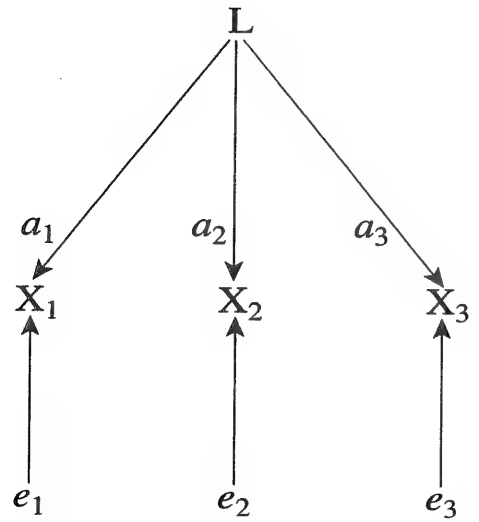

Figure 2.6 A diagram of a parallel tests model, in which all pathways from the latent variable (L) to the items (X_1, X_2, X_3) are equal in value to one another, as are all pathways from the error terms to the items

proportion of variance shared between two variables. So, if correlations between the true score and each of two items are equal, the proportions of variance shared between the true score and each item also must be equal. Assume that a true score contributes the same *amount* of variance to each of two items. This amount can be an equal *proportion* of total variance for each item only if the items have identical total variances. In order for the total variances to be equal for the two items, the amount of variance each item receives from sources other than the true score must also be equal. As all variation sources other than the true score are lumped together as error, this means that the two items must have equal error variances. For example, if X_1 got 9 arbitrary units of variation from its true score and 1 from error, the true score proportion would be 90% of total variation. If X_2 also got 9 units of variation from the true score, these 9 units could only be 90% of the total if the total variation were 10. The total could only equal 10 if error contributed 1 unit to X_2, as it did to X_1. The correlation between each item and the true score then would equal the square root of the proportion of each item's variance that is attributable to the true score, or roughly .95 in this case.

Thus, because the parallel tests model assumes that the amount of influence from the latent variable is the same for each item *and* that the amount from other sources (error) is the same for each item, the proportions of item variance attributable to the latent variable and to error are equal for all items. This also means that, under the assumptions of parallel tests, standardized path coefficients from the latent variable to each item are equal for all items. It was assuming that standardized path coefficients were equal that made it possible, in an earlier example, to compute path coefficients from correlations between items. The path diagram rule, discussed earlier, relating path coefficients to correlations, should help us to understand why these equalities hold when one accepts the preceding assumptions.

The assumptions of this model also imply that correlations among items are identical (e.g., the correlation between X_1 and X_2 is identical to the correlation between X_1 and X_3 or X_2 and X_3). How do we arrive at this conclusion from the assumptions? The correlations are all the same because the only mechanism to account for the correlation between any two items is the route through the latent variable that links those items. For example, X_1 and X_2 are linked only by the route made up of paths a_1 and a_2. The correlation can be computed by tracing the route joining the two items in question and multiplying the path values. For any two items, this entails multiplying two paths that have identical values (i.e., $a_1 = a_2 = a_3$). Correlations computed by multiplying equal values will, of course, be equal.

The assumptions also imply that each of these correlations between items equals the square of any path from the latent variable to an individual item. How do we reach this conclusion? The product of two different paths (e.g., a_1 and a_2) is identical to the square of either path because both path coefficients are identical. If $a_1 = a_2 = a_3$, and $(a_1 \times a_2) = (a_1 \times a_3) = (a_2 \times a_3)$, then each of these latter products must also equal the value of any of the a-paths multiplied by itself.

It also follows from the assumptions of this model that the proportion of error associated with each item is the complement of the proportion of variance that is related to the latent variable. In other words, whatever effect on a given item is not explained by the latent variable must be explained by error. Together, these two effects explain 100% of the variation in any given item. This is so simply because the error term, e, is defined as encompassing all sources of variation in the item other than the latent variable.

These assumptions support at least one other conclusion: Because each item is influenced equally by the latent variable and each error term's influence on its corresponding item is also equal, the items all have equal means and equal variances. If the only two sources that can influence the mean are identical for all items, then clearly the means for the items also will be identical. This reasoning also holds for the item variances.

In conclusion, the parallel tests model assumes

1. random error
2. errors are not correlated with each other
3. errors are not correlated with the true score
4. the latent variable affects all items equally
5. the amount of error for each item is equal

These assumptions allow us to reach a variety of interesting conclusions. Furthermore, the model enables us to make inferences about the latent variable, based on the items' correlations with one another. However, the model accomplishes this feat by setting forth fairly stringent assumptions.

ALTERNATIVE MODELS

As it happens, all of the narrowly restrictive assumptions associated with strictly parallel tests are not necessary in order to make useful inferences about the relationship of true scores to observed scores. A model based on what are technically called *essentially tau-equivalent tests* (or, occasionally, *randomly parallel tests*) makes a more liberal assumption, namely, that the amount of error variance associated with a given item need not equal the error variance of the other items (e.g., Allen & Yen, 1979). Consequently, the *standardized* values of the paths from the latent variable to each item may not be equal. However, the *unstandardized* values of the path from the latent variable to each item (i.e., the *amount* as opposed to *proportion* of influence that the latent variable has on each item) are still presumed to be identical for all items. This means that items are parallel with respect to how much they are influenced by the latent variable but are not necessarily influenced to exactly the same extent by extraneous factors that are lumped together as error. Under strictly parallel assumptions, different items not only tap the true score to the same degree, but their error components are also the same. Tau equivalency ("tau" is the Greek equivalent of "t," as in true score) is much easier to live with because it does not impose the "equal errors" condition. Because errors may vary, item means and variances may also vary. The more liberal assumptions of this model are attractive because finding equivalent measures of equal variance is rare. This model allows us to reach many of the same conclusions we reach with strictly parallel tests but with less restrictive assumptions. Readers may wish to compare this model to Nunnally and Bernstein's (1994) discussion of the "domain sampling model."

Some scale developers consider even the essentially tau-equivalent model too restrictive. After all, how often can we assume that each item is influenced by the latent variable to the same degree? Tests developed under what is called the *congeneric model* (Jöreskog, 1971) are subject to an even more relaxed set of assumptions (see Carmines & McIver, 1981, for a discussion of congeneric tests). It assumes (beyond the basic measurement assumptions) merely that all the items share a common latent variable. They need not bear equally strong relationships to the latent variable, and their error variances need not be equal. One must assume only that each item reflects the true score to some degree. Of course, the more strongly each item correlates with the true score, the more reliable the scale will be.

An even less constrained approach is the *general factor model,* which allows multiple latent variables to underlie a given set of items. Carmines and McIver (1981), Loehlin (1998), and Long (1983) have discussed the merits of this type of very general model, chief among them being its improved correspondence to real-world data. Structural equation modeling approaches often incorporate factor analyses into their measurement models. Situations in which multiple latent variables underlie a set of indicators exemplify the general factor model (Loehlin, 1998).

The congeneric model is a special case of the factor model (i.e., a single-factor case). Likewise, an essentially tau-equivalent measure is a special case of a congeneric measure—one for which the relationships of items to their latent variable are assumed to be equal. Finally, a strictly parallel test is a special case of an essentially tau-equivalent one, adding the assumption of equal relationships between each item and its associated sources of error.

Another measurement strategy should be mentioned. This is *item response theory* (IRT). This approach has been used primarily, but not exclusively, with dichotomous-response (e.g., correct versus incorrect) items in developing ability tests. Different models within the broader class of IRTs may be based on the normal or, with increasing frequency, the logistic probability function. IRT assumes that each individual item has its own characteristic sensitivity to the latent variable, represented by an item-characteristic curve (ICC). An ICC is a plot of the relationship between the value of the latent variable (e.g., ability) and the probability of a certain response to an item (e.g., answering it correctly). Thus the curve reveals how much ability an item demands to be answered correctly. We will consider IRT further in Chapter 7.

Except for that consideration of IRT in Chapter 7 and a discussion of factor analysis in Chapter 6, we will focus primarily on parallel and essentially tau-equivalent models for several reasons. First, they exemplify "classical" measurement theory. In addition, discussing the mechanisms by which other models operate can quickly become burdensome. Finally, classical models

have proven very useful for social scientists with primary interests other than measurement who, nonetheless, take careful measurement seriously. This group is the audience for whom the present text has been written. For these individuals, the scale development procedures that follow from a classical model generally yield very satisfactory scales. Indeed, although to my knowledge no tally is readily available, I suspect that (outside of ability testing) a substantial majority of the well-known and highly regarded scales used in social science research were developed using such procedures.

EXERCISES

1. How can we infer the relationship between the latent variable and two items related to it, based on the correlations between the two items?
2. What is the chief difference in assumptions between the parallel tests and essentially tau-equivalent models?
3. Which measurement model assumes, beyond the basic assumptions common to all measurement approaches, only that the items share a common latent variable?

NOTE

1. Although -.70 is also an allowable square root of .49, deciding between the positive or negative root is typically of less concern than one would think. As long as all the items can be made to correlate positively with one another (if necessary, by "reverse scoring" certain items as discussed in Chapter 5), then the signs of the path coefficients from the latent variable to the individual items will be the same and are arbitrary. Note, however, that giving positive signs to these paths implies that the items indicate more of the construct, whereas negative coefficients would imply the opposite.

3

Reliability

Reliability is a fundamental issue in psychological measurement. Its importance is clear once its meaning is fully understood. Scale *reliability* is the proportion of variance attributable to the true score of the latent variable. There are several methods for computing reliability, but they all share this fundamental definition. However, how one conceptualizes and operationalizes reliability differs based on the computational method one uses.

CONTINUOUS VERSUS DICHOTOMOUS ITEMS

Although items may have a variety of response formats, we assume in this chapter that item responses consist of multiple-value response options. Dichotomous items (i.e., items with only two response options, such as "yes" and "no" or with multiple response options that can be classified as "right" versus "wrong") are widely used in ability testing and, to a lesser degree, in other measurement contexts. Examples are

1. Zurich is the capital of Switzerland.	True	False	
2. What is the value of pi?	(a) 1.41	(b) 3.14	(c) 2.78

Special methods for computing reliability that take advantage of the computational simplicity of dichotomous responses have been developed. General measurement texts such as Nunnally and Bernstein (1994) cover these methods in some detail. The logic of these methods for assessing reliability largely parallels the more general approach that applies to multipoint, continuous scale items. In the interest of brevity, this chapter will make only passing reference to reliability assessment methods intended for scales made up of dichotomous items. Some characteristics of this type of scale are discussed in Chapter 5.

INTERNAL CONSISTENCY

Internal consistency reliability, as the name implies, is concerned with the homogeneity of the items within a scale. Scales based on classical measurement

models are intended to measure a single phenomenon. As we saw in the preceding chapter, measurement theory suggests that the relationships among items are logically connected to the relationships of items to the latent variable. If the items of a scale have a strong relationship to their latent variable, they will have a strong relationship to each other. Although we cannot directly observe the linkage between items and the latent variable, we can certainly determine whether the items are correlated to one another. A scale is internally consistent to the extent that its items are highly intercorrelated. What can account for correlations among items? There are two possibilities: that items causally effect each other (e.g., item A causes item B) or that the items share a common cause. Under most conditions, the former explanation is unlikely, leaving the latter as the more obvious choice. High interitem correlations thus suggest that the items are all measuring (i.e., are manifestations of) the same thing. If we make the assumptions discussed in the preceding chapter, we also can conclude that strong correlations among items imply strong links between items and the latent variable. Thus, a unidimensional scale or a single dimension of a multidimensional scale should consist of a set of items that correlate well with each other. Multidimensional scales measuring several phenomena—for example, the Multidimensional Health Locus of Control (MHLC) scales (Wallston et al., 1978)—are really families of related scales; each "dimension" is a scale in its own right.

Coefficient Alpha

Internal consistency is typically equated with Cronbach's (1951) coefficient alpha, α. We will examine alpha in some detail for several reasons. First, it is widely used as a measure of reliability. Second, its connection to the definition of reliability may be less self-evident than is the case for other measures of reliability (such as the alternate forms methods) discussed later. Consequently, alpha may appear more mysterious than other reliability computation methods to those who are not familiar with its internal workings. Finally, an exploration of the logic underlying the computation of alpha provides a sound basis for comparing how other computational methods capture the essence of what is meant by reliability.

The Kuder-Richardson formula 20, or KR-20, as it is more commonly known, is a special version of alpha for items that are dichotomous (e.g., Nunnally & Bernstein, 1994). However, as noted earlier, we will concentrate on the more general form that applies to items having multiple response options.

You can think about all the variability in a set of item scores as due to one of two things: (a) actual variation across individuals in the phenomenon that the

scale measures (i.e., true variation in the latent variable) and (b) error. This is true because classical measurement models define *the phenomenon* (e.g., patients' desire for control of their interactions with a physician) as the source of all shared variation and *error* as any remaining, or unshared, variation in scale scores (e.g., a single item's unintended double meaning). Another way to think about this is to regard total variation as having two components: *signal* (i.e., true differences in patients' desire for control) and *noise* (i.e., score differences caused by everything but true differences in desire for control). Computing alpha, as we shall see, partitions the total variance among the set of items into signal and noise components. The proportion of total variation that is signal equals alpha. Thus another way to think about alpha is that it equals 1 – error variance, or, conversely, that error variance = 1 – alpha.

The Covariance Matrix

To understand internal consistency more fully, it helps to examine the covariance matrix of a set of scale items. A covariance matrix for a set of scale items reveals important information about the scale as a whole.

A covariance matrix is a more general form of a correlation matrix. In a correlation matrix, the data have been standardized, with the variances set to 1.0. In a covariance matrix, the data entries are unstandardized; thus, it contains the same information, in unstandardized form, as a correlation matrix. The diagonal elements of a covariance matrix are variances—covariances of items with themselves—just as the unities along the main diagonal of a correlation matrix are variables' variances standardized to 1.0 and also their correlations with themselves. Its off-diagonal values are covariances, expressing relationships between pairs of unstandardized variables just as correlation coefficients do with standardization. So, conceptually, a covariance matrix consists of (a) variances (on the diagonal) for individual variables and (b) covariances (off-diagonal) representing the unstandardized relationship between variable pairs.

A typical covariance matrix for three variables X_1, X_2, and X_3 is shown in Table 3.1.

An alternative that somewhat more compactly uses the customary symbols for matrices, variances, and covariances is

$$\begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} \\ \sigma_{1,2} & \sigma_2^2 & \sigma_{2,3} \\ \sigma_{1,3} & \sigma_{2,3} & \sigma_3^2 \end{bmatrix}$$

TABLE 3.1
Variances and Covariances for Three Variables

	X_1	X_2	X_3
X_1	Var_1	$Cov_{1,2}$	$Cov_{1,3}$
X_2	$Cov_{1,2}$	Var_2	$Cov_{2,3}$
X_3	$Cov_{1,3}$	$Cov_{2,3}$	Var_3

Covariance Matrices for Multi-Item Scales

Let us focus our attention on the properties of a covariance matrix for a set of items that, when added together, make up a scale. The covariance matrix presented above has three variables, X_1, X_2, and X_3. Assume that these variables are actually scores for three items and that the items, X_1, X_2, and X_3, when added together make up a scale we will call Y. What can this matrix tell us about the relationship of the individual items to the scale as a whole?

A covariance matrix has a number of very interesting (well, useful, at least) properties. Among these is the fact that adding all of the elements in the matrix together (i.e., summing the variances, which are along the diagonal, and the covariances, which are off of the diagonal) gives a value that is exactly equal to the variance of the scale as a whole, assuming that the items are equally weighted. So, if we add all the terms in the symbolic covariance matrix, the resulting sum would be the variance of scale Y. This is very important and bears repeating: The variance of a scale, Y, made up of any number of items, equals the sum of all the values in the covariance matrix for those items, assuming equal item weighting.[1] Thus the variance of a scale Y, made up of three equally weighted items, X_1, X_2, and X_3, has the following relationship to the covariance matrix of the items: $\sigma_y^2 = C$, where

$$C = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} \\ \sigma_{1,2} & \sigma_2^2 & \sigma_{2,3} \\ \sigma_{1,3} & \sigma_{2,3} & \sigma_3^2 \end{bmatrix}$$

Readers who would like more information about the topics covered in this section are referred to Nunnally (1978) for covariance matrices and Namboodiri (1984) for an introduction to matrix algebra in statistics. The

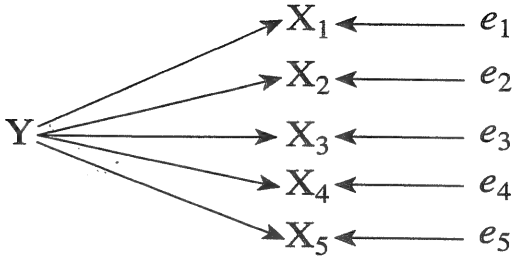

Figure 3.1 Diagrammatic representation of how a set of five items relates to the common latent variable Y

covariance matrix for the individual items has additional useful information not mentioned here. Applications that can be derived from item covariance matrices are discussed by Bohrnstedt (1969).

Alpha and the Covariance Matrix

Alpha is defined as the proportion of a scale's total variance that is attributable to a common source, presumably the true score of a latent variable underlying the items. Thus if we want to compute alpha, it would be useful to have a value for the scale's total variance and a value for the proportion that is "common" variance. The covariance matrix is just what we need in order to do this.

Recall the diagram we used in Chapter 2 to show how items related to their latent variable, as in Figure 3.1.

All of the variation in items that is due to the latent variable, Y, is shared or common. (The terms *joint* and *communal* are also used to describe this variation.) When Y varies (as it will, for example, across individuals having different levels of the attribute it represents), scores on all the items will vary with it because it is a cause of those scores. Thus if Y is high, all the item scores will tend to be high; if Y is low, they will tend to be low. This means that the items will tend to vary jointly (i.e., be correlated with one another). So, the latent variable affects all of the items and thus they are correlated. The error terms, in contrast, are the source of the unique variation that each item possesses. Whereas all items share variability due to Y, no two share any variation from the same error source, under our classical measurement assumptions. The value of a given error term only affects the score of one item. Thus, the error terms

are not correlated with one another. So, each item (and, by implication the scale defined by the sum of the items) varies as a function of (a) the source of variation common to itself and the other items and (b) unique, unshared variation that we refer to as error. It follows that the total variance for each item, and hence for the scale as a whole, must be a combination of variance from common and unique sources. According to the definition of *reliability*, alpha should equal the ratio of common-source variation to total variation.

Now, consider a k-item measure called Y whose covariance matrix is as follows:

$$\begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} & \cdots & \sigma_{1,k} \\ \sigma_{1,2} & \sigma_2^2 & \sigma_{2,3} & \cdots & \sigma_{2,k} \\ \sigma_{1,3} & \sigma_{2,3} & \sigma_3^2 & \cdots & \sigma_{3,k} \\ \vdots & \vdots & \vdots & & \vdots \\ \sigma_{1,k} & \sigma_{2,k} & \sigma_{3,k} & \cdots & \sigma_k^2 \end{bmatrix}$$

The variance, σ_y^2, of the k-item scale equals the sum of all matrix elements. The entries along the main diagonal are the variances of the individual items represented in the matrix. The variance of the ith item is signified as σ_i^2. Therefore, the sum of the elements along the main diagonal, $\sum\sigma_i^2$, is the sum of the variances of the individual items. Thus the covariance matrix gives us ready access to two values: (a) the total variance of the scale, σ_y^2, defined as the sum of all elements in the matrix and (b) the sum of the individual item variances, $\sum\sigma_i^2$, computed by summing entries along the main diagonal. These two values can be given a conceptual interpretation. The sum of the whole matrix is, by definition, the variance of Y, the scale made up of the individual items. However, this total variance, as we have said, can be partitioned into different parts.

Let us consider how the covariance matrix separates common from unique variance by examining how the elements on the main diagonal of the covariance matrix differ from all the off-diagonal elements. All of the variances (diagonal elements) are single-variable or "variable-with-itself" terms. I noted earlier that these variances can be thought of as covariances of items with themselves. Each variance contains information about only one item. In other

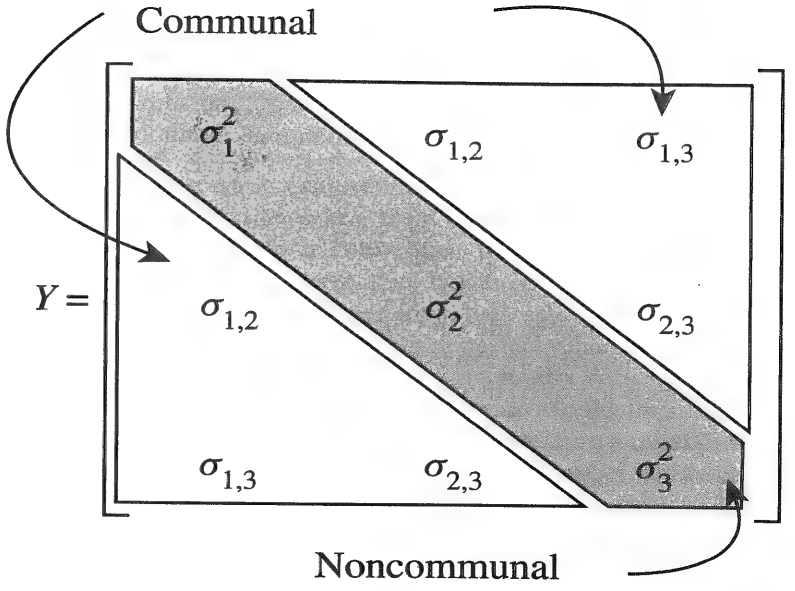

Figure 3.2 A variance-covariance matrix showing that the variances along the main diagonal (shaded area) are noncommunal, whereas the covariances lying above and below the diagonal (unshaded area) are communal

words, each represents information that is based on a single item, not joint variation shared among items. (Within that single item, some of its variation will be due to the common underlying variable and thus shared with other items; some will not. However, the item's variance does not quantify the extent of shared variance, merely the amount of dispersion in the scores for that item, irrespective of what causes it.) The off-diagonal elements of the covariance matrix all involve pairs of terms, and thus common, or joint, variation between two of the scale's items (covariation). Thus the elements in the covariance matrix (and hence the total variance of Y) consist of covariation (*joint variation*, if you will) plus *nonjoint*, or *noncommunal*, variation concerning items considered individually. Figure 3.2 pictorially represents these two sub-divisions of the covariance matrix. The shaded area along the diagonal is the noncommunal portion of the matrix, and the two off-diagonal regions within the triangular borders are, together, the communal portion.

As the covariances—and only the covariances—represent communal variation, all noncommunal variation must be represented in the variances along the main diagonal of the covariance matrix and thus by the term $\sum\sigma_i^2$. The total variance, of course, is expressed by σ_y^2, the sum of all the matrix elements. Thus we can express the ratio of nonjoint variation to total variation in Y as

$$\sum\sigma_i^2 / \sigma_y^2$$

This ratio corresponds to the sum of the diagonal values in the covariance matrix. It thus follows that we can express the proportion of joint, or communal, variation as what is left over, that is, the complement of this value as shown:

$$1 - \left(\Sigma\sigma_i^2 \Big/ \sigma_y^2\right)$$

This value corresponds to the sum of all the off-diagonal values of the covariance matrix. It may seem strange, or at least inefficient, to compute the diagonal elements and then subtract them from the value of the covariance matrix as a whole. Why not just compute the sum of the off-diagonal elements directly as $\sum\sigma_{i,j}$, where i and j represent each of the two items involved in a particular covariance? In fact, one would arrive at the same exact point by directly computing the sum of off-diagonal elements. The formula involving subtraction from 1 is a legacy of the days when computers were not available to do calculations. Computing the total variance for Y and the variance for each individual item, i, were probably operations that had already been done for other purposes. Even if there were no need to calculate these variances for other purposes, consider the computational effort involved. For a 20-item scale, the choice would be between computing 21 variances (one for each item and another for the entire scale) and 190 covariances (i.e., one for each of the 380 off-diagonal elements of the matrix, with those above the diagonal identical to those below) plus the total variance. Thus, a formula that quantifies communal variance as what remains after removing noncommunal from total variance makes more sense than might at first be apparent.

The value represented by the formula

$$1 - \left(\Sigma\sigma_i^2 \Big/ \sigma_y^2\right)$$

or, equivalently,

$$\Sigma\sigma_i^2 \Big/ \sigma_y^2$$

would at first blush seem to capture the definition of alpha, that is, the communal portion of total variance in a scale that can be attributed to the items' common source, which we presume reflects the true score of the latent variable. We still need one more correction, however. This need becomes apparent if we consider what would happen if we had, say, five perfectly correlated items. Such an arrangement should result in perfect reliability. The correlation matrix in this instance would consist of a 5×5 matrix with all values equal to 1.0. The denominator of the preceding equation would thus equal 25. The numerator, however, would only equal 20, thus yielding a reliability of 20/25, or .80 rather than 1.0. Why is this so? The total number of elements in the covariance matrix is k^2. The number of elements in the matrix that are noncommunal (i.e., those along the main diagonal) is k. The number that are communal (all those not on the diagonal) is $k^2 - k$. The fraction in our last formula thus has a numerator based on $k^2 - k$ values and a denominator based on k^2 values. To adjust our calculations so that the ratio expresses the relative magnitudes rather than the numbers of terms that are summed in the numerator and denominator, we multiply the entire expression representing the proportion of communal variation by values to counteract the differences in numbers of terms summed. To do this, we multiply by $k^2/(k^2 - k)$, or, equivalently, $k/(k - 1)$. This limits the range of possible values for alpha to between 0.0 and 1.0. In the five-item example just discussed, multiplying .80 by 5/4 yields the appropriate 1.0. Readers may want to do the mental arithmetic for matrices of other sizes. It should soon become apparent that $k/(k - 1)$ is always the multiplier that will yield an alpha of 1.0 when the items are all perfectly correlated. Thus, we arrive at the usual formula for coefficient alpha:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\Sigma\sigma_i^2}{\sigma_{y_i}^2}\right)$$

To summarize, a measure's reliability equals the proportion of total variance among its items that is due to the latent variable and thus is communal. The formula for alpha expresses this by specifying the portion of total variance for the item set that is unique, subtracting this from 1 to determine the

proportion that is communal, and multiplying by a correction factor to adjust for the number of elements contributing to the earlier computations.

An Alternative Formula for Alpha

Another common formula for computing alpha is based on correlations rather than covariances. Actually, it uses $\bar{r}$, the average interitem correlation. This formula is

$$\alpha = \frac{k\bar{r}}{1 + (k-1)\bar{r}}$$

It follows logically from the covariance-based formula for alpha. Consider the covariance formula in conceptual terms:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\text{Sum of item variances}}{\text{Sum of variances and covariances}}\right)$$

Note that the numerator and denominator in the term on the right are sums of individual values. However, the sum of these individual values is identical to the mean of the values multiplied by the number of values involved. (For example, k numbers that sum to 50 and k times the mean of those numbers both equal 50. To illustrate further, substitute 10 for k in the preceding sentence; the average of 10 values that sum to 50 has to be 5, and 10 times 5 equals 50, the same value as the original sum.) Therefore, the numerator of the term on the right must equal k times the average item variance, $\bar{v}$, and the denominator must equal k times the average variance plus $(k^2 - k)$—or, alternatively, $(k)(k-1)$—times the average covariance, $(\bar{c})$:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{k\bar{v}}{k\bar{v} + (k)(k-1)\bar{c}}\right)$$

To remove the "1" from the equation, we can replace it with its equivalent $[k\bar{v} + (k)(k-1)\bar{c}]/[k\bar{v} + (k)(k-1)\bar{c}]$, which allows us to consolidate the whole term on the right into a single ratio:

$$\alpha = \frac{k}{k-1}\left(\frac{k\bar{v} + k(k-1)\bar{c} - k\bar{v}}{k\bar{v} + (k)(k-1)\bar{c}}\right)$$

or, equivalently,

$$\alpha = \frac{k}{k-1}\left(\frac{k(k-1)\bar{c}}{k[\bar{v} + (k-1)\bar{c}]}\right)$$

Cross-canceling k from the numerator of the left term and denominator of the right term, while cross-canceling $(k-1)$ from the numerator of the right term and the denominator of the left term, yields the simplified expression:

$$\alpha = \frac{k\bar{c}}{\bar{v} + (k-1)\bar{c}}$$

Recall that the formula we are striving for involves correlations rather than covariances and thus standardized rather than unstandardized terms. After standardizing, an average of covariances is identical to an average of correlations, and a variance equals 1.0. Consequently, we can replace $\bar{c}$ with the average interitem correlation, $\bar{r}$, and $\bar{v}$ with 1.0. This yields the correlation-based formula for coefficient alpha:

$$\alpha = \frac{k\bar{r}}{1 + (k-1)\bar{r}}$$

This formula is known as the Spearman-Brown prophecy formula, and one of its important uses will be illustrated in the section of this chapter dealing with split-half reliability computation.

The two different formulas, one based on covariances and the other on correlations, are sometimes referred to as the raw score and standardized score formulas for alpha, respectively. The *raw score* formula preserves information about item means and variances in the computation process, because covariances are based on values that retain the original scaling of the raw data. If items have markedly different variances, those with larger variances will be given greater weight than those with lesser variances when this formula is used to compute alpha. The *standardized score* formula based on correlations does not retain the original scaling metric of the items. Recall that a correlation is a standardized covariance. So, all items are placed on a common metric and thus weighted equally in the computation of alpha by the standardized formula. Which is better depends on the specific context and whether equal weighting is desired. As we shall see in later chapters, recommended

procedures for developing items often entail structuring their wording so as to yield comparable variances for each item. When these procedures are followed, there is typically little difference in the alpha coefficients computed by the two alternative methods. On the other hand, when procedures aimed at producing equivalent item variances are not followed, observing that the standardized and raw alpha values differ appreciably (e.g., by .05 or more) is indicative that at least one item has a variance that differs appreciably from the variances of the other items.

Reliability and Statistical Power

An often overlooked benefit of more reliable scales is that they increase *statistical power* for a given sample size (or allow a smaller sample size to yield equivalent power), relative to less reliable measures. To have a specified degree of confidence in the ability to detect a difference of a given magnitude between two experimental groups, for example, one needs a particular size sample. The probability of detecting such a difference (i.e., the power of the statistical test) can be increased by increasing the sample size. In many applications, much the same effect can be obtained by improving the reliability of measurement. A reliable measure, like a larger sample, contributes relatively less error to the statistical analysis. Researchers might do well to weigh the relative advantages of increasing scale reliability versus sample size in research situations where both options are available.

The power gains from improving reliability depend on a number of factors, including the initial sample size, the probability level set for detecting a Type I error, the effect size (e.g., mean difference) that is considered significant, and the proportion of error variance that is attributable to measure unreliability rather than sample heterogeneity or other sources. A precise comparison between reliability enhancement and sample size increase requires that these factors be specified; however, the following examples illustrate the point. In a hypothetical research situation with the probability of a Type I error set at .01, a 10-point difference between two means regarded as important, and an error variance equal to 100, the sample size would have to be increased from 128 to 172 (a 34% increase) to raise the power of an F test from .80 to .90. Reducing the total error variance from 100 to 75 (a 25% decrease) would have essentially the same result without increasing the sample size. Substituting a highly reliable scale for a substantially poorer one might accomplish this. As another example, for $N = 50$, two scales with reliabilities of .38 that have a correlation ($r = .24$) barely achieving significance at $p < .10$ are significant at $p < .01$ if their reliabilities are increased to .90. If the reliabilities remained at .38, a sample more than twice as large would be needed for the correlation to reach

$p < .01$. Lipsey (1990) provides a more comprehensive discussion of statistical power, including the effects of measurement reliability.

RELIABILITY BASED ON CORRELATIONS BETWEEN SCALE SCORES

There are alternatives to internal consistency reliability. These types of reliability computation involve having the same set of people complete two separate versions of a scale or the same version on multiple occasions.

Alternate Forms of Reliability

If two strictly parallel forms of a scale exist, then the correlation between them can be computed as long as the same people complete both parallel forms. For example, assume that a researcher first developed two equivalent sets of items measuring patients' desire for control when interacting with physicians, then administered both sets of items to a group of patients and, finally, correlated the scores from one set of items with the scores from the other set. This correlation would be the alternate forms reliability. Recall that parallel forms are made up of items, all of which (either within or between forms) do an equally good job of measuring the latent variable. This implies that both forms of the scale have identical alphas, means, and variances and measure the same phenomenon. In essence, parallel forms consist of one set of items that has been divided more or less arbitrarily into two subsets that make up the two parallel, alternate forms of the scale. Under these conditions, the correlation between one form and the other is equivalent to correlating either form with itself, as each alternate form is equivalent to the other.

Split-Half Reliability

A problem with alternate forms reliability is that we usually do not have two versions of a scale that conform strictly to the assumptions of parallel tests. However, there are other reliability estimates that apply the same sort of logic to a single set of items. Because alternate forms are essentially made up of a single pool of items that has been divided in two, it follows that we can (a) take the set of items that makes up a single scale (i.e., a scale that does not have any alternate form), (b) divide that set of items into two subsets, and (c) correlate the subsets to assess reliability.

A reliability measure of this type is called a *split-half reliability*. Split-half reliability is really a class rather than a single type of computational method because there are a variety of ways in which the scale can be split in half. One method is to compare the first half of the items to the second half. This type of *first-half last-half split* may be problematic, however, because factors other than the value of the latent variable (in other words, sources of error) might affect each subset differently. For example, if the items making up the scale in question were scattered throughout a lengthy questionnaire, the respondents might be more fatigued when completing the second half of the scale. Fatigue would then differ systematically between the two halves and would thus make them appear less similar. However, the dissimilarity would not be so much a characteristic of the items per se as of their position in the item order of the scale. Other factors that might differentiate earlier occurring from later occurring items are a practice effect whereby respondents might get better at answering the items as they go along, failure to complete the entire set of items, and possibly even something as mundane as changes in the print quality of a questionnaire from front to back. As with fatigue, these factors would lower the correlation between halves because of the order in which the scale items were presented and not because of the quality of the scale items. As a result of factors such as these, measuring the strength of the relationships among items may be complicated by circumstances not directly related to item quality, resulting in an erroneous reliability assessment.

To avoid some of the pitfalls associated with item order, one can assess another type of split-half reliability known as *odd-even reliability*. In this instance, the subset of odd-numbered items is compared to the even-numbered items. This ensures that each of the two subsets of items consists of an equal number from each section (i.e., the beginning, middle, and end) of the original scale. Assuming that item order is irrelevant (as opposed to the "easy-to-hard" order common to achievement tests, for example), this method avoids many of the problems associated with first-half versus second-half split halves.

In theory, there are many other ways to arrive at split-half reliability. Two alternatives to the methods discussed above for constituting the item subsets are *balanced halves* and *random halves*. In the former case, one would identify some potentially important item characteristics (such as first-person wording, item length, or whether a certain type of response indicates the presence or absence of the attribute in question). The two halves of the scale would then be constituted so as to have the characteristics equally represented in each half. Thus an investigator might divide up the items so that each subset had the same number of items worded in the first-person, the same number of short items, and so on. However, when considering multiple item characteristics, it might be impossible to balance the proportion of one without making it impossible to

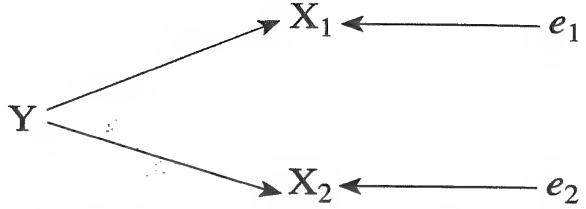

Figure 3.3 A path diagram showing the relationship of two split halves of a measure (X_1 and X_2) to their common latent variable

balance another. This would be the case, for example, if there were more long than short first-person items. Creating a balance for the latter characteristic would necessitate an imbalance of the former. Also, it may be difficult to determine which characteristics of the items should be balanced.

An investigator could obtain random halves merely by randomly allocating each item to one of the two subsets that will eventually be correlated with one another to compute the reliability estimate. How well this works depends on the number of items, the number of characteristics of concern, and the degree of independence among the characteristics. Hoping that a small number of items, varying along several interrelated dimensions, will yield comparable groupings through randomization is unrealistic. On the other hand, randomly assigning a set of 50 items varying with respect to two or three uncorrelated characteristics to two categories might yield reasonably comparable subsets.

Which method of achieving split halves is best depends on the particular situation. What is most important is that the investigator think about how dividing the items might result in nonequivalent subsets and what steps can be taken to avoid this. The reasoning behind both split-halves and alternate-forms reliability is a natural extension of the parallel tests model.

Although when we initially discussed that model, we regarded each item as a "test," one can also regard a scale (or the two halves of a scale) that conforms to the model as a "test." Therefore, we can apply the logic we used in the case of several items to the case of two alternate forms or two halves of a scale. Consider two "tests" (scale halves or alternate forms) under the parallel tests assumptions.

The only route linking the two consists of the causal paths from the latent variable to each. Thus the product of these paths' values equals the correlation between the tests. If the path values have to be equal (and they do, under the

assumptions of this model), then the correlation between the tests equals the square of the path value from latent variable to either test. The square of that path (assuming that it is a standardized path coefficient) is also the proportion of variance in either test that is influenced by the latent variable. This, in turn, is the definition of reliability. Thus the correlation between the two tests equals the reliability of each.

Whereas the "tests" referred to in the preceding paragraph are two complete versions of a scale in the alternate forms case, they are two half-scales in the split-half instance. Thus the correlation between two split-halves yields a reliability estimate for each *half* of the whole set of items, which is an underestimate of the reliability for the entire set. An estimate of the reliability of the entire scale, based on the reliability of a portion of the scale, can be computed by using the Spearman-Brown formula, discussed earlier in this chapter. Recall that, according to this formula,

$$\alpha = \frac{k\bar{r}}{1 + (k-1)\bar{r}}$$

where k is the number of items in question and $\bar{r}$ is the average correlation of any one item with any other (i.e., the average interitem correlation). If you had determined the reliability of a subset of items (e.g., by means of the split-half method) and knew how many items that reliability was based on (e.g., half the number in the whole scale), you could use the formula to compute $\bar{r}$. Then, you could plug that value of $\bar{r}$ and the number of items in the *whole* scale back into the formula. The result would be an estimate of the reliability of the whole scale, based on a reliability value computed on split halves of the scale. It simplifies matters if you perform a little algebra on the Spearman-Brown equation to put it into the following form:

$$\bar{r} = \frac{r_{yy}}{[k - (k-1)r_{yy}]}$$

where r_{yy} is the reliability of the item set in question. For example, if you knew that the split-half reliability for two 9-item halves was equal to .90, you could compute $\bar{r}$ as follows:

$$\bar{r} = \frac{.9}{[9 - (8)(.9)]} = .5$$

You could then recompute the reliability for the whole 18-item scale by using $\bar{r} = .5$ and $k = 18$ in the Spearman-Brown formula. Thus, the reliability estimate for the full scale is

$$\frac{18 \times .5}{1 + (17 \times .5)}$$

which equals 9/9.5 or .947. (Note that increasing the number of items has increased the reliability. A quick look at the Spearman-Brown formula should make it apparent that, all else being equal, a longer scale will always be more reliable than a shorter one.)

Temporal Stability

Another two-score method of computing reliability involves the *temporal stability* of a measure, or how constant scores remain from one occasion to another. *Test-retest reliability* is the method typically used to assess this. Suppose that, instead of developing two sets of items to measure patients' desire for control when interacting with physicians, our hypothetical investigator developed only a single set. Those items could be given to one group of patients on two separate occasions, and the scores from the first occasion could be correlated with those from the later administration. The rationale underlying reliability determinations of this type is that if a measure truly reflects some meaningful construct, it should assess that construct comparably on separate occasions. In other words, the true score of the latent variable should exert comparable influence on observed scores on two (or more) occasions, while the error component should not remain constant across administrations of the scale. Consequently, the correlation of scores obtained across two administrations of a scale to the same individuals should represent the extent to which the latent variable determines observed scores. This is the equivalent to the definition of reliability as the proportion of variance attributable to the true score of the latent variable.

The problem with this reasoning is that what happens to the scores over time may or may not have to do with the error-proneness of the measurement procedure. Nunnally (1978) points out that characteristics of the items might cause them to yield temporally stable responses even when the construct of interest has changed. For example, if a purported anxiety measure was influenced by social desirability as well as anxiety, scores might remain constant despite variations in anxiety. The stability in scores, reflected in a high correlation across occasions of administration, would not be the result of invariance

in the phenomenon of interest. Alternatively, the phenomenon may not change while scores on the measure do; that is, the scale could be unreliable. Or, changes in scores may be attributed to unreliability when, in fact, the phenomenon itself has changed and the measure has accurately tracked that change. The problem is that either a change or the absence of a change can be due to a variety of things besides the (un)reliability of the measurement procedure. Kelly and McGrath (1988) identified four factors that are confounded when one examines two sets of scores on the same measure, separated in time. These are (a) real change in the construct of interest (e.g., a net increase in average level of anxiety among a sample of individuals), (b) systematic oscillations in the phenomenon (e.g., variations in anxiety, around some constant mean, as a function of time of day), (c) changes attributable to differences in subjects or measurement methods rather than the phenomenon of interest (e.g., *fatigue* effects that cause items to be misread), and (d) temporal instability due to the inherent unreliability of the measurement procedure. Of these factors, only the fourth is unreliability. These authors also note that, although methods such as the multitrait-multimethod matrix approach (discussed in the next chapter) can help, it is never possible to unconfound these factors fully.

This is not to say that demonstrating temporal stability is unimportant. In any number of research contexts, it may be critical to assume (or demonstrate) that measurements separated in time are highly correlated. However, the stability we seek in these situations is stability of both the measure and the phenomenon. Test-retest correlations only tell us about the measure when we are highly confident that the phenomenon has remained stable. Such confidence is not often warranted. Thus test-retest reliability, although important, may best be thought of as revealing something about the nature of a phenomenon and its measurement, not the latter alone. Referring to invariance in scores over time as *temporal stability* is preferable because it does not suggest, as does test-retest reliability, that measurement error is the source of any instability we observe.

GENERALIZABILITY THEORY

Thus far, our discussion of reliability has focused on partitioning observed variation into the portion that is attributable to the true score of the latent variable and the remaining portion, which is error. This section briefly introduces a more general framework for partitioning variance among error and nonerror sources.

Before we apply the idea of a finer partitioning of error variance to measurement, let us consider a more general research example in which multiple

sources of variation are examined. Suppose that a researcher wanted to determine the effectiveness of a training program intended to increase professional productivity. Assume, furthermore, that the researcher administered the training program to a large sample of college professors and to a comparable sample of artists. The researcher also identified comparable groups of professors and artists who would not participate in the training program but would take part in the same productivity assessment as the training program participants. Upon giving the study some thought, this researcher might have concluded that the observations of productivity would reflect the operation of three identifiable sources of systematic variation: (a) participant versus nonparticipant, (b) professor versus artist, and (c) the interaction of these effects. A reasonable analytic strategy in this situation would be to perform an analysis of variance (ANOVA) on the productivity scores, treating each of these sources of variation as a dimension in the analysis. The investigator could then determine to what extent each source of variation contributed to the total variation in professional productivity. In essence, this analytic strategy would partition the total variance among observed productivity scores into several sources: training participation, profession, the interaction of these, and error. Error would represent all sources of variation other than those specified by the preceding factors.

Now, consider a hypothetical situation in which a researcher is developing a scale of desire for autonomy. The measure will be used in a study of elderly people, some of whom may have visual problems. Consequently, the investigator plans to administer the desire-for-autonomy measure orally to those people who would have difficulty reading, and in written form to the remaining study participants.

If the researcher ignored mode of administration (written versus oral) as a source of variation in test scores, he or she would be regarding each score obtained as due to the true level of the respondent's desire for autonomy plus some degree of error. The researcher could proceed to calculate reliability as discussed earlier. Note, however, that merely computing alpha on the scale scores without regard for the mode of administration would not differentiate the potential systematic error due to administration method from any other source of error.

Alternatively, it is possible for the researcher to acknowledge administration mode as a source of variation among scores, using an analysis of variance approach. If the resulting analysis demonstrated that the difference between administration methods accounted for an inconsequential proportion of the total variation in scores, then the researcher could have greater confidence in the comparability of scores for individuals completing either the oral or written version. If, on the other hand, a significant amount of the total observed

variation in scores were attributable to administration mode, then the researcher would know that any interpretation of scores should take this difference between modes into consideration.

Generalizability theory (e.g., Cronbach, Gleser, Nanda, & Rajaratnam, 1972) provides a framework for examining the extent to which one can assume equivalence of a measurement process across one or more dimensions. In the preceding example, the dimension in question was mode of administration. Each dimension of interest is a potential source of variation and is referred to as a *facet*. The example focused on mode of administration as the only potential source of variation (other than individuals) across which the investigator wished to generalize. Therefore, this example involves a single facet.

In the parlance of generalizability theory, observations obtainable across all levels of a facet (e.g., with both oral and written administration of the scale) constitute a universe of admissible observations. The mean of these observations is referred to as the *universe score* and is analogous to the true score of classical test theory (Allen & Yen, 1979). A study aimed at determining to what extent scores are comparable across different levels of a facet is called a *generalizability study*, or *G-study*. The hypothetical study of desire for autonomy is an example of a G-study by virtue of its addressing the effects of different "levels" of the mode-of-administration facet.

The purpose of the G-study is to help the investigator determine the extent to which the facet does or does not limit generalizability. If a facet (e.g., mode of administration) explains a significant amount of the variance in observed scores, findings *do not* generalize across levels (e.g., oral versus written administration) of that facet. The extent to which one can generalize across levels of the facet without misrepresenting the data is expressed as a *generalizability coefficient*. This is typically computed by forming a ratio from the appropriate mean squares resulting from the ANOVA performed as part of the G-study. Conceptually, the generalizability coefficient is the ratio of universe score variance to observed score variance and is analogous to the reliability coefficient (Allen & Yen, 1979). Note, however, that if a G-study yields a poor generalizability coefficient, the study's design points to a source of the problem—that is, the facet examined. A reliability coefficient merely identifies the amount of error without attributing it to a specific source.

In some instances, choosing the appropriate ANOVA design, deciding which effects correspond to the facets of interest, and constructing the correct generalizability coefficient can be demanding. Just as with analysis of variance in general, multiple dimensions, nested, crossed, and mixed effects can complicate a G-study. (See Myers, 1979, or Kirk, 1995, for general discussions of ANOVA designs.) Keeping the design of a G-study simple is advisable. It is also prudent to consult a source that explains in detail how to build

the appropriate ANOVA model for a given type of G-study. Crocker and Algina (1986) describe the appropriate designs for several different one- and two-facet generalizability studies. This source also provides a good general introduction to generalizability theory.

SUMMARY

Scales are reliable to the extent that they consist of reliable items that share a common latent variable. Coefficient alpha corresponds closely to the classical definition of reliability as the proportion of variance in a scale that is attributable to the true score of the latent variable. Various methods for computing reliability have different utility in particular situations. For example, if one does not have access to parallel versions of a scale, computing alternate forms reliability is impossible. A researcher who understands the advantages and disadvantages of alternative methods for computing reliability is in a good position to make informed judgments when designing a measurement study or evaluating a published report.

EXERCISES[2]

1. If a set of items has good internal consistency, what does that imply about the relationship of the items to their latent variable?
2. In this exercise,[3] assume that the following is a covariance matrix for a scale, Y, made up of three items, X_1, X_2, and X_3:

1.2	.5	.4
.5	1.0	.6
.4	.6	1.8

   (a) What are the variances of X_1, X_2, and X_3?

   (b) What is the variance of Y?

   (c) What is coefficient alpha for scale Y?

3. Discuss the ways in which test-retest reliability confounds other factors with the actual scale properties.
4. How does the logic of alternate forms reliability follow from the assumptions of parallel tests?

NOTES

1. For weighted items, covariances are multiplied by products and variances by squares of their corresponding item weights. See Nunnally (1978, pp. 154-156) for a more complete description of this.

2. Throughout the book, the solution for any exercise that requires a numeric answer will be found in the Notes section of the chapter in which the exercise appears.

3. The answers are (a) 1.2, 1.0, and 1.8 (which sum to 4.0); (b) 7.0 (the sum of all elements in the matrix); (c) $(3/2) \times [1 - (4.0/7.0)] = 0.64$.

4

Validity

Whereas *reliability* concerns how much a variable influences a set of items, *validity* concerns whether the variable is the underlying cause of item covariation. To the extent that a scale is reliable, variation in scale scores can be attributed to the true score of some phenomenon that exerts a causal influence over all the items. However, determining that a scale is reliable does not guarantee that the latent variable shared by the items is, in fact, the variable of interest to the scale developer. The adequacy of a scale as a measure of a *specific variable* (e.g., perceived psychological stress) is an issue of validity.

Some authors have assigned a broader meaning to validity. For example, Messick (1995) described six types of validity, one of which (consequential validity) concerns the impact on respondents of how their scores are used. Although Messick's (1995) views on validity have raised some thought-provoking issues, his classification system has not been widely adopted. According to the more conventional interpretation, validity is inferred from the manner in which a scale was constructed, its ability to predict specific events, or its relationship to measures of other constructs. There are essentially three types of validity that correspond to these operations:

1. content validity
2. criterion-related validity
3. construct validity

Each type will be reviewed briefly. For a more extensive treatment of validity, including a discussion of methodological and statistical issues in criterion-related validity and alternative validity indices, see Ghiselli, Campbell, and Zedeck (1981, Chapter 10). Readers might also want to consider Messick's (1995) more all-encompassing view of validity.

CONTENT VALIDITY

Content validity concerns item sampling adequacy—that is, the extent to which a specific set of items reflects a content domain. Content validity is

easiest to evaluate when the domain (e.g., all the vocabulary words taught to sixth graders) is well defined. The issue is more subtle when measuring attributes, such as beliefs, attitudes, or dispositions, because it is difficult to determine exactly what the range of potential items is and when a sample of items is representative. In theory, a scale has content validity when its items are a randomly chosen subset of the universe of appropriate items. In the vocabulary test example used above, this is easily accomplished. All the words taught during the school year would be defined as the universe of items. Some subset could then be sampled. However, in the case of measuring beliefs, for example, we do not have a convenient listing of the relevant universe of items. Still, one's methods in developing a scale (e.g., having items reviewed by experts for relevance to the domain of interest, as suggested in Chapter 5) can help to maximize item appropriateness. For example, if a researcher needed to develop a measure contrasting expected outcomes and desired outcomes (e.g., expecting versus wanting a physician to involve the patient in decision making), it might be desirable for her or him to establish that all relevant outcomes were represented in the items. To do this, the researcher might ask colleagues familiar with the context of the research to review an initial list of items and suggest content areas that have been omitted but should be included. Items reflecting this content could then be added.

CRITERION-RELATED VALIDITY

In order to have *criterion-related validity*, as the term implies, an item or scale is required to have only an empirical association with some criterion or "gold standard." Whether or not the theoretical basis for that association is understood is irrelevant to criterion-related validity. If one could show, for example, that dowsing is empirically associated with locating underground water sources, then dowsing would have validity with respect to the criterion of successful well digging. Thus criterion-related validity per se is more of a practical issue than a scientific one, because it is concerned not with understanding a process but merely with predicting it. In fact, criterion-related validity is often referred to as *predictive validity.*

Criterion-related validity by any name does not necessarily imply a causal relationship among variables, even when the time ordering of the predictor and the criterion are unambiguous. Of course, prediction in the context of theory (e.g., prediction as a hypothesis) *may* be relevant to the causal relationships among variables and can serve a very useful scientific purpose.

Another point worth noting about criterion-related validity is that, logically, one is dealing with the same type of validity issue whether the criterion follows, precedes, or coincides with the measurement in question. Thus, in addition to "predictive validity," *concurrent validity* (e.g., "predicting" driving skill from answers to oral questions asked during the driving test) or even *postdictive validity* (e.g., "predicting" birth weight from an infancy developmental status scale) may be used more or less synonymously with criterion-related validity. The most important aspect of criterion-related validity is not the time relationship between the measure in question and the criterion whose value one is attempting to infer but, rather, the strength of the empirical relationship between the two events. The term *criterion-related validity* has the advantage over the other terms of being temporally neutral and thus is preferable.

Criterion-Related Validity Versus Accuracy

Before leaving criterion-related validity, a few words are in order concerning its relationship to accuracy. As Ghiselli and colleagues (1981) point out, the correlation coefficient, which has been the traditional index of criterion-related validity, may not be very useful when predictive accuracy is the issue. A correlation coefficient, for example, does not reveal how many cases are correctly classified by a predictor (although tables that provide an estimate of the proportion of cases falling into various percentile categories, based on the size of the correlation between predictor and criterion, are described by Ghiselli et al., 1981, p. 311). It may be more appropriate in some situations to divide both a predictor and its criterion into discrete categories and to assess the "hit rate" for placing cases into the correct category of the criterion based on their predictor category. For example, one could classify each variable into "low" versus "high" categories, and conceptualize accuracy as the proportion of correct classifications (i.e., instances when the value of the predictor corresponds to the value of the criterion). Where one divides categories is an important consideration. Consider a criterion that has two nonarbitrary states, such as "sick" and "well, " and an assessment tool that has a range of scores that an investigator wants to dichotomize. The purpose of the assessment tool is to predict whether people will test as positive or negative for the sickness in question. Because the outcome is dichotomous, it makes sense to make the predictor dichotomous. There are two possible errors in classification: the measure can mistakenly classify a truly sick person as well (false negative) or a truly well person as sick (false positive). Where along the range of scores on the assessment tool the dividing line is placed when dichotomizing can affect the rates of these two types of errors. At the extremes, classifying everyone as well will avoid any false negatives (but increase false positives), whereas

classifying everyone as sick will avoid any false positives (but increase false negatives). Obviously, in both of these extreme cases, the assessment tool would have no predictive value at all. The goal, of course, is to choose a cutoff that produces the fewest errors of either type, and thus the highest accuracy. Often, there is no ideal cut point, that is, one resulting in perfect classification. In such a case, the investigator may make a conscious effort to minimize one type of error rather than the other. For example, if the sickness is devastating and the treatment is effective, inexpensive, and benign, the cost of a false negative (resulting in undertreating) is far greater than the cost of a false positive (resulting in overtreating). Thus, choosing a cutoff so as to reduce false negatives while accepting false positives would seem appropriate. On the other hand, if the remedy is both expensive and unpleasant and the sickness mild, the opposite trade-off might make more sense.

Also, it is important to remember that, even if the correlation between a predictor measure and a criterion is perfect, the score obtained on the predictor is not an estimate of the criterion. Correlation coefficients are insensitive to linear transformations of one or both variables. A high correlation between two variables implies that scores on those variables obtained from the same individual will occupy similar locations on their respective distributions. For example, someone scoring very high on the first variable is likely also to score very high on the second, if the two are strongly correlated. *Very high,* however, is a relative rather than an absolute term and does not consider the two variables' units of measurement, for example. Transforming the predictor's units of measurement to that of the criterion may be necessary to obtain an accurate numerical prediction. This adjustment is equivalent to determining the appropriate intercept in addition to the slope of a regression line. A failure to recognize the need to transform a score could lead to erroneous conclusions. An error of this sort is perhaps most likely to occur if the predictor happens to be calibrated in units that fall into the same range as the criterion. Assume, for example, that someone devised the following "speeding ticket scale" to predict how many tickets drivers would receive over 5 years.

1. I exceed the speed limit when I drive.

   Frequently *Occasionally* *Rarely* *Never*

2. On multilane roads, I drive in the passing lane.

   Frequently *Occasionally* *Rarely* *Never*

3. I judge for myself what driving speed is appropriate.

   Frequently *Occasionally* *Rarely* *Never*

Let us also make the implausible assumption that the scale correlates perfectly with the number of tickets received in a 5-year period. The scale is scored by giving each item a value of 3 when a respondent circles "frequently," 2 for "occasionally," 1 for "rarely," and 0 for "never." The item scores then are summed to get a scale score. The score's perfect criterion validity does not mean that a score of 9 translates into nine tickets over 5 years. Rather, it means that the people who score highest on the instrument are also the people who have the highest observed number of tickets in a year. Some empirically determined transformation (e.g., .33 × SCORE) would yield the actual estimate. This particular transformation would predict three tickets for a driver scoring 9. If criterion-related validity were high, then a more accurate estimate could be computed. However, the similarity between the numerical values of the criterion and the predictor measure prior to an appropriate transformation would have nothing to do with the degree of validity.

CONSTRUCT VALIDITY

Construct validity (Cronbach & Meehl, 1955) is directly concerned with the theoretical relationship of a variable (e.g., a score on some scale) to other variables. It is the extent to which a measure "behaves" the way that the construct it purports to measure should behave with regard to established measures of other constructs. So, for example, if we view some variable, based on theory, as positively related to constructs A and B, negatively related to C and D, and unrelated to X and Y, then a scale that purports to measure that construct should bear a similar relationship to measures of those constructs. In other words, our measure should be positively correlated with measures of constructs A and B, negatively correlated with measures of C and D, and uncorrelated with measures of X and Y. A depiction of these hypothesized relationships might look like Figure 4.1.

The extent to which empirical correlations match the predicted pattern provides some evidence of how well the measure "behaves" like the variable it is supposed to measure.

Differentiating Construct From Criterion-Related Validity

People often confuse construct and criterion-related validity because the same exact correlation can serve either purpose. The difference resides more in the investigator's intent than in the value obtained. For example, an epidemiologist might attempt to determine which of a variety of measures

	A	B	C	D	X	Y
Variable	+	+	−	−	0	0

Figure 4.1 A hypothesized relationship among variables

obtained in a survey study correlate with health status. The intent might be merely to identify risk factors without concern (at least initially) for the underlying causal mechanisms linking scores on measures to health status. Validity, in this case, is the degree to which the scales can predict health status. Alternatively, the concern could be more theoretical and explanatory. The investigator, like the epidemiologist described in this book's opening chapter, might endorse a theoretical model that views stress as a cause of health status, and the issue might be how well a newly developed scale measures stress. This might be assessed by evaluating the "behavior" of the scale relative to how theory suggests stress should operate. If the theory suggested that stress and health status should be correlated, then the same empirical relationship used as evidence of predictive validity in the preceding example might be used as evidence of construct validity.

So-called known-groups validation is another example of a procedure that can be classified as either construct or criterion-related validity, depending on the investigator's intent. *Known-groups validation* typically involves demonstrating that some scale can differentiate members of one group from another, based on their scale scores. The purpose may be either theory related (such as when a measure of attitudes toward a certain group is validated by correctly differentiating those who do or do not affiliate with members of that group) or purely predictive (such as when one uses a series of seemingly unrelated items to predict job turnover). In the former case, the procedure should be considered a type of construct validity and in the latter, criterion-related validity.

How Strong Should Correlations Be in Order to Demonstrate Construct Validity?

There is no cutoff that defines construct validity. It is important to recognize that two measures may share more than construct similarity. Specifically, similarities in the way that constructs are measured may account for some covariation in scores, independent of construct similarity. For example, two variables scored on a multipoint scoring system (with scores from 1 to 100)

will have a higher correlation with each other than with a binary variable, all else being equal. This is an artifact caused by the structure of the measurement methods. Likewise, because of procedural similarities, data of one type gathered by interviews may correlate to a degree with other data gathered in the same way; that is, some of the covariation between two variables may be due to measurement similarity rather than construct similarity. This fact provides some basis for answering the question concerning the magnitude of correlations necessary to conclude construct validity. The variables, at a minimum, should demonstrate covariation above and beyond what can be attributed to shared method variance.

Multitrait-Multimethod Matrix

Campbell and Fiske (1959) devised a procedure called the multitrait-multimethod matrix that is extremely useful for examining construct validity. The procedure involves measuring more than one construct by means of more than one method so that one obtains a "fully crossed" method-by-measure matrix. For example, suppose that a study is designed in which anxiety, depression, and shoe size are each measured at two times using two different measurement procedures each time. (Note that two different samples of individuals could be measured at the same time. What effect would this have on the logic of the approach?) Each construct could be assessed by two methods, a visual analog scale (a line upon which respondents make a mark to indicate the amount of the attribute they possess, be it anxiety, depression, or bigness of foot) and a rating assigned by an interviewer following a 15-minute interaction with each subject. One could then construct a matrix of correlations obtained between measurements as in Table 4.1.

Another possible distinction, not in the table, is between related versus unrelated traits. Because the entries that reflect the same trait (construct) and the same method should share both method and construct variance, one would expect these correlations to be highest. It is hoped that correlations corresponding to the same trait but different methods would be the next highest. If so, this would suggest that construct covariation is higher than method covariation; in other words, our measures were more influenced by what was measured than by how it was measured. In contrast, there is no reason why any covariation should exist between shoe size and either of the other two constructs when they are measured by different procedures. Thus these correlations should not be significantly different from zero. For nonidentical but theoretically related constructs, such as depression and anxiety, one would expect some construct covariation. This is potentially a highly informative

TABLE 4.1
Interpretations of Correlations in a Multitrait-Multimethod Matrix

Time 2 \ Time 1	A_v	A_i	D_v	D_i	S_v	S_i
A_v	TM	T	M		M	
A_i	T	TM		M		M
D_v	M		TM	T	M	
D_i		M	T	TM		M
S_v	M		M		TM	T
S_i		M		M	T	TM

NOTE: TM stands for same trait and method (reliability); T stands for same trait, different method; M stands for same method, different trait; *A, D,* and *S* refer to the constructs anxiety, depression, and shoe size; the subscripts *v* and *i* refer to visual analog and interview methods.

set of correlations for establishing construct validity. If, for example, our depression measures were both well established but our anxiety measures were currently being developed, we could assess the amount of covariation attributable to concept similarity under conditions of similar and different measurement procedures. Theory asserts that anxiety and depression should be substantially correlated even when measured by different methods. If this proved to be the case, it would serve as evidence of the construct validity of our new anxiety measures. More specifically, these correlations would be indicative of *convergent validity,* evidence of similarity between measures of theoretically related constructs. Ideally, the correlations between anxiety and depression would be less than those between two depression or two anxiety measures, but they would be substantially greater than between either of the depression scores and shoe size. Equally important is evidence that the anxiety measures did not correlate significantly with measures of shoe size, irrespective of similarity or dissimilarity of measurement technique. This is evidence of *discriminant validity* (sometimes called *divergent validity*), which is the absence of correlation between measures of unrelated constructs. Shoe size and anxiety correlating significantly when measured the same way would suggest that method per se accounted for a substantial amount of the variation (and covariation) associated with similar measures of the dissimilar constructs.

Mitchell (1979) observed that the methods involved in collecting data for a multitrait-multimethod matrix constitute a two-facet G-study (see Chapter 3), with traits and methods being the facets. The multitrait-multimethod matrix allows us to partition covariation into "method" and "trait" (or "construct") sources. We can thus make more precise statements about construct validity because it allows us to differentiate covariation that truly reflects similarity of construct (and thus is relevant to construct validity) from covariation that is an artifact of applying similar measurement procedures (and thus does not relate to construct validity). Such a differentiation is not possible when one simply examines a single correlation between two measures.

WHAT ABOUT FACE VALIDITY?

Many people use the term *face validity* to describe a set of items that assess what they appear to measure on their face. In my view, this usage is unfortunate for several reasons.

First, the assumption that a measure assesses what it looks like it is assessing can be wrong. For example, Idler and Benyamini (1997) examined 27 large, well-conducted epidemiological studies to determine precisely what a very common item was tapping. That item asks people to evaluate their overall health as "poor," "fair," "good," or "excellent." Most people would judge this single-item measure to assess exactly what it says: the respondent's health. Idler and Benyamini noted that the item was an excellent predictor of a variety of health outcomes. It consistently outperformed other variables in accounting for variance across the different studies. More relevant to our discussion, it appeared not to be related primarily to health status. Models often contained the single item and also established measures of health status. Typically, both the single-item health self-rating and the other health status measures were significant predictors in the same model. That is, they did not share sufficient variance for the predictive contribution of one to preclude an independent predictive contribution from the other. Rather, the single-item health self-rating appeared to share variance to a greater degree with psychological variables. These findings suggest that this extensively used single item is *not* a valid indicator of health status, as it appears on its face. For this item, looking like it is measuring what we want is not enough to support claims of validity.

A second problem with evaluating a measure based on face validity is that there are times when it is important that the variable being measured is not evident. For example, an instrument intended to assess the degree to which

people answer untruthfully (e.g., to make themselves "look good") would hardly benefit from having its purpose apparent to respondents. Would we conclude that it was invalid because it did not look as if it were measuring untruthfulness? Hopefully not. So, here we have a case where failure to look like what it actually is cannot support a conclusion of invalidity.

A final concern with face validity is that it is unclear to whom an instrument's purpose should be evident, on its face. Is it the respondent? If a physician asks a patient if he or she has been more thirsty than usual, is the validity of that question dependent upon the patient knowing why it was asked? Clearly not. Is it the person creating the instrument who should recognize the purpose? It is hard to imagine that the linkage between instrument content and variable of interest is not obvious to an instrument's creator (except, perhaps, in cases concerning purely empirical, atheoretical criterion-related validity). If this meaning of face validity were adopted, essentially all scales would be judged valid. Finally, is it a broader scientific community that should recognize an instrument's purpose based on its appearance? This interpretation is very likely to yield conflicting evidence. An item that looks like it measures one variable to some experts might look like it measures another to a second, equally qualified group. Often, it seems, people who claim that a scale is or is not valid because it does or does not appear to have face validity are basing their claim on personal perceptions. That is, if the intent and appearance of an instrument look similar *to them*, they are inclined to consider it face valid; otherwise, they reject the idea that it is face valid. This seems like a feeble basis for any claim of validity.

Depending on the circumstances, there may be advantages or disadvantages to an instrument's intent being evident from its appearance. As we shall see in the next chapter, the item-generation process often produces statements that refer explicitly to the variable of interest. This usually is not a bad thing. I am not suggesting that instruments generally should be constructed so that their intent is not evident from appearances. Rather, I am suggesting that whether or not that is the case has little or nothing to do with validity.

EXERCISES

1. Give an example of how the same correlation between a scale and a behavior might be indicative of either construct validity or criterion-related validity. Explain how both (a) the motives behind computing the correlation and (b) the interpretation of that correlation would differ, depending on the type of validity the investigator was trying to assess.

2. Assume that an investigator has paper-and-pencil measures of two constructs: self-esteem and social conformity. The investigator also has interview-based scores on the same two constructs. How could these data be used in a multitrait-multimethod matrix to demonstrate that the method of data collection had an undesirably strong effect on the results obtained?

5

Guidelines in Scale Development

Thus far, the material presented has been fairly abstract. We now look at how this knowledge can be applied. This chapter provides a set of specific guidelines that investigators can use in developing measurement scales.

STEP 1: DETERMINE CLEARLY WHAT IT IS YOU *WANT* TO MEASURE

This is deceptively obvious, and many researchers *think* they have a clear idea of what they wish to measure, only to find that their ideas are more vague than they thought. Frequently, this realization occurs after considerable effort has been invested in generating items and collecting data—a time when changes are far more costly than if discovered at the outset of the process. Should the scale be based in theory, or should you strike out in new intellectual directions? How specific should the measure be? Should some aspect of the phenomenon be emphasized more than others?

Theory as an Aid to Clarity

As noted in Chapter 1, thinking clearly about the content of a scale requires thinking clearly about the construct being measured. Although there are many technical aspects involved in developing and validating a scale, one should not overlook the importance of being well grounded in the substantive theories related to the phenomenon to be measured. The types of scales that are the primary focus of this book are intended to measure elusive phenomena that cannot be observed directly. Because there is no tangible criterion against which one can compare this type of scale's performance, it is important to have some clear ideas to serve as a guide. The boundaries of the phenomenon must be recognized so that the content of the scale does not inadvertently drift into unintended domains.

Theory is a great aid to clarity. Relevant social science theories should always be considered before developing a scale of the type discussed in this volume. If it turns out that extant theory offers no guide to the scale

developers, then they may decide that a new intellectual direction is necessary. However, this decision should be an informed one, reached only after reviewing appropriate theory related to the measurement problem at hand. Even if there is no available theory to guide the investigators, they must lay out their own conceptual formulations prior to trying to operationalize them. In essence, they must specify at least a tentative theoretical model that will serve as a guide to scale development. This may be as simple as a well-formulated definition of the phenomenon they seek to measure. Better still would be to include a description of how the new construct relates to existing phenomena and their operationalizations.

Specificity as an Aid to Clarity

The level of specificity or generality at which a construct is measured also may be very important. There is general agreement in the social sciences that variables will relate most strongly to one another when they match with respect to level of specificity (see Ajzen & Fishbein, 1980, for a discussion of this). Sometimes a scale is intended to relate to very specific behaviors or constructs, whereas at other times a more general and global measure is sought.

As an illustration of measures that differ in specificity, consider the locus of control construct. *Locus of control* (LOC) is a widely used concept that concerns individuals' perceptions about who or what influences important outcomes in their lives. The construct can be applied broadly, as a means of explaining patterns of global behavior spanning many situations, or narrowly, to predict how an individual will respond in a very specific context. The sources of influence also can be described either broadly or specifically. Rotter's (1966) Internal-External (I-E) scale, for example, is concerned at a fairly general level with these perceptions. A single dimension ranging from personal control to control by outside factors underlies the scale, and the outcomes on which the items focus are general, such as personal success. The external sources of control also are described in general terms. The following external statement is from Rotter's I-E scale: "The world is run by the few people in power, and there is not much the little guy can do about it." Levenson (1973) developed a multidimensional LOC scale that allows for three loci of control: oneself, powerful other people, and chance or fate. This permits an investigator to look at external sources of control a bit more specifically by characterizing them as either powerful others or fate. The outcomes on which she focused, however, remained general. An example of an item from Levenson's Powerful Others subscale is, "I feel like what happens in my life is determined by powerful others." Wallston, Wallston, and DeVellis

(1978) developed the Multidimensional Health Locus of Control (MHLC) scales using Levenson's three loci of control, with outcomes specific to health, such as avoiding illness or getting sick. A sample item from the Powerful Others scale of the MHLC is: "Having regular contact with my physician is the best way for me to avoid illness." Wallston, Stein, and Smith (1994) subsequently developed an even more outcome-specific health locus of control measure (MHLC form C) that consists of a series of "template" items. This measure allows the researcher to specify any health problem of interest by substituting the name of the illness or disorder for the phrase "my condition" in each of the template items. A sample item from the Powerful Others scale of MHLC Form C, as it might be used in a study of diabetes, is: "If I see my doctor regularly, I am less likely to have problems with my diabetes."

Each of these progressively more specific LOC scales is potentially useful. Which is most useful depends largely upon what level of outcome or locus generality relates to the scientific question being asked. For example, if a locus of control scale is intended to predict a general class of behavior or will be compared to other variables assessing constructs at a general level, then Rotter's scale may be the best choice because it, too, is general. On the other hand, if a researcher is interested in predicting specifically how beliefs about the influence of other people affects certain health behaviors, then the Wallston, Stein and Smith (1994) scale may be more appropriate because the level of specificity matches that research question. During its development, each of these scales had a clear frame of reference that determined what level of specificity was appropriate, given the intended function of the scale. The point is that scale developers should make this determination as an active decision and not merely generate a set of items and then see what it looks like after the fact.

The locus of control example illustrated specificity with respect to outcomes (e.g., how the world is run versus problems with diabetes) and the loci of control (i.e., external generally versus fate and powerful others separately). However, scale specificity can vary along a number of dimensions, including content domains (e.g., anxiety versus psychological adjustment more broadly), setting (e.g., questionnaires designed specifically for relevance to particular work environments), and population (e.g., children versus adults or military personnel versus college students).

Being Clear About What to Include in a Measure

Scale developers should ask themselves if the construct they wish to measure is distinct from other constructs. As noted earlier, scales can be developed

to be relatively broad or narrow with respect to the situations to which they apply. This is also the case with respect to the constructs they cover. Measuring general anxiety is perfectly legitimate. Such a measure might assess both test anxiety and social anxiety. This is fine if it matches the goals of the scale developer or user. However, if one is interested in only one specific type of anxiety, then the scale should exclude all others. Items that might "cross over" into a related construct (e.g., tapping social anxiety when the topic of interest is test anxiety) can be problematic.

Sometimes, apparently similar items may tap quite different constructs. In such cases, although the purpose of the scale may be to measure one phenomenon, it may also be sensitive to other phenomena. For example, certain depression measures, such as the Center for Epidemiological Studies Depression (CES-D) scale (Radloff, 1977), have some items that tap somatic aspects of depression (e.g., concerning the respondent's ability to "get going"). In the context of some health conditions, such as arthritis, these items might mistake aspects of the illness for symptoms of depression (see Blalock, DeVellis, Brown, & Wallston, 1989, for a discussion of this specific point). A researcher developing a new depression scale might choose to avoid somatic items if the scale were to be used with certain populations (e.g., the chronically ill) or with other measures of somatic constructs (such as hypochondriasis). Used for other purposes, of course, it might be very important to include somatic items, as when the line of investigation specifically concerns somatic aspects of negative affect.

STEP 2: GENERATE AN ITEM POOL

Once the purpose of a scale has been clearly articulated, the developer is ready to begin constructing the instrument in earnest. The first step is to generate a large pool of items that are candidates for eventual inclusion in the scale.

Choose Items That Reflect the Scale's Purpose

Obviously, these items should be selected or created with the specific measurement goal in mind. The description of exactly what the scale is intended to do should guide this process. Recall that all items making up a homogeneous scale should reflect the latent variable underlying them. Each item can be thought of as a "test," in its own right, of the strength of the latent variable. Therefore, the content of each item should primarily reflect the construct of interest. Multiple items will constitute a more reliable test than

individual items, but each must still be sensitive to the true score of the latent variable.

Theoretically, a good set of items is chosen randomly from the universe of items relating to the construct of interest. The universe of items is assumed to be infinitely large, which pretty much precludes any hope of actually identifying it and extracting items randomly. However, this ideal should be kept in mind. If you are writing items anew, as is so often the case, you should think creatively about the construct you seek to measure. What other ways can an item be worded so as to get at the construct? Although the items should not venture beyond the bounds of the defining construct, they should exhaust the possibilities for types of items within those bounds. The properties of a scale are determined by the items that make it up. If they are a poor reflection of the concept you have worked long and hard to articulate, then the scale will not accurately capture the essence of the construct.

It is also important that the "thing" that items have in common is truly a construct and not merely a category. Recall, once again, that our models for scale development regard items as overt manifestations of a common latent variable that is their cause. Scores on items related to a common construct are determined by the true score of that construct. However, as noted in Chapter 1, just because items relate to a common category does not guarantee that they have the same underlying latent variable. Such terms as *attitudes*, *barriers to compliance*, or *life events* often define categories of constructs rather than the constructs themselves. A pool of items that will eventually be the basis of a unidimensional scale should not merely share a focus on attitudes, for example, but on specific attitudes, such as attitudes toward punishing drug abusers. One can presumably envision a characteristic of the person, a latent variable, if you will, that would "cause" responses to items dealing with punishing drug abusers. It is quite a challenge to imagine a characteristic that accounts for attitudes in general. The same is true for the other examples cited. Barriers to compliance are typically of many types. Each type (e.g., fear of discovering symptoms, concern over treatment costs, anticipation of pain, distance of treatment facilities, perceptions of invulnerability) may represent a latent variable. There may even be nontrivial correlations among some of the latent variables. However, each of these barriers is a separate construct. Thus the term *barriers* describes a category of constructs rather than an individual construct related to a single latent variable. Items measuring different constructs that fall within the same category (e.g., perceptions of invulnerability and concerns over treatment costs) should not be expected to covary the way items do when they are manifestations of a common latent variable.

Redundancy

At this stage of the scale development process, it is better to be overinclusive, all other things being equal. Redundancy is *not* a bad thing when developing a scale. In fact, the theoretical models that guide our scale development efforts are based on redundancy. In discussing the Spearman-Brown Prophecy formula in Chapter 3, I pointed out that reliability varies as a function of the number of items, all else being equal. We are attempting to capture the phenomenon of interest by developing a set of items that reveals the phenomenon in different ways. By using multiple and seemingly redundant items, the content that is common to the items will summate across items while their irrelevant idiosyncracies will cancel out. Without redundancy, this would be impossible. Useful redundancy pertains to the construct, not incidental aspects of the items. Changing nothing more than an "a" to "the" in an item will certainly give you redundancy with respect to the important content of the item, but it will also be redundant with respect to many things that you want to vary, such as the basic grammatical structure and choice of words. On the other hand, two items, such as "I will do almost anything to ensure my child's success" and "No sacrifice is too great if it helps my child achieve success," may be usefully redundant because they express a similar idea in somewhat different ways.

You can tolerate considerably more redundancy in your item pool than in the final scale, even though some redundancy is desirable even in the latter. For example, if you have an item such as "In my opinion, pet lovers are kind," there is obviously little advantage in including an additional item that states "In my estimation, pet lovers are kind." These items clearly tap similar sentiments regarding pet ownership, but they also share a common grammatical structure and use nearly identical vocabularies. However, an item such as "I think that people who like pets are good people" might do a good job of being redundant with respect to the substantive content of the first item—without trivial redundancy. At this very early stage of scale development, however, even the extreme redundancy of the first two items in this example might be acceptable, as long as only one appears on the final scale. Considering two items, even when they are as similar as these, might provide the scale developer with an opportunity to compare them and express a preference (e.g., that "opinion" may seem less pretentious than "estimation"). This opportunity would be lost if only one of the two items were considered.

Number of Items

It is impossible to specify the number of items that should be included in an initial pool. Suffice it to say that you want considerably more than you plan to

include in the final scale. Recall that internal consistency reliability is a function of how strongly the items correlate with one another (and hence with the latent variable) and how many items you have in the scale. As the nature of the correlations among items is usually not known at this stage of scale development, having lots of items is a form of insurance against poor internal consistency. The more items you have in your pool, the fussier you can be about choosing ones that will do the job you intend. It would not be unusual to begin with a pool of items that is three or four times as large as the final scale. Thus a 10-item scale might evolve from a 40-item pool. If items are particularly difficult to generate for a given content area or if empirical data indicate that numerous items are not needed to attain good internal consistency, then the initial pool may be as small as 50% larger than the final scale.

In general, the larger the item pool, the better. However, it is certainly possible to develop a pool too large to administer on a single occasion to any one group of subjects. If the pool is exceptionally large, the researcher can eliminate some items based on a priori criteria, such as lack of clarity, questionable relevance, or undesirable similarity to other items.

Beginning the Process of Writing Items

Getting started writing items is often the most difficult part of the item-generation process. Let me describe how I begin this process. At this point, I am less interested in item quality than in merely expressing relevant ideas. I often begin with a statement that is a paraphrase of the construct I want to measure. For example, if I were interested in developing a measure of self-perceived susceptibility to commercial messages, I might begin with the statement "I am susceptible to commercial messages." I then would try to generate additional statements that get at the same idea somewhat differently. My next statement might be, "Commercial messages affect me a lot." I would continue in this manner, imposing virtually no quality standards on the statements. My goal at this early stage is simply to identify a wide variety of ways in which the central concept of the intended instrument can be stated. As I write, I may seek alternative ways of expressing critical ideas. For example, I might substitute "the things that I see in TV or magazine ads" for "commercial messages" in the next set of sentences. I find that writing quickly and uncritically is useful. After generating perhaps three or four times the number of items that I anticipate including in the final instrument, I look over what I have written. Now is the time to become critical. Items can be examined for how well they capture the central ideas and for clarity of expression. The sections that follow delineate some of the specific item characteristics to avoid or incorporate in the process of selecting from and revising the original statement list.

Characteristics of Good and Bad Items

Listing all the things that make an item good or bad is an impossible task. The content domain obviously has a significant bearing on item quality. However, there are some characteristics that reliably separate better items from worse ones. Most of these relate to clarity. As pointed out in Chapter 1, a good item should be unambiguous. Questions that leave the respondent in a quandary should be eliminated.

Scale developers should avoid *exceptionally lengthy items,* as length usually increases complexity and diminishes clarity. However, it is not desirable to sacrifice the meaning of an item in the interest of brevity. If a modifying clause is essential to convey the intent of an item, then include it. However, avoid unnecessary wordiness. In general, an item such as "I often have difficulty making a point" will be better than an unnecessarily longer one, such as "It is fair to say that one of the things I seem to have a problem with much of the time is getting my point across to other people."

Another related consideration in choosing or developing items is the *reading difficulty level* at which the items are written. There are a variety of methods (e.g., Dale & Chall, 1948; Fry, 1977) for assigning grade levels to passages of prose, including scale items. These typically equate longer words and sentences with higher reading levels. Reading most local newspapers presumably requires a sixth-grade reading level.

Fry (1977) delineates several steps to quantifying reading level. The first is to select a sample of text that begins with the first word of a sentence and contains exactly 100 words. (For scales having only a few items, you may have to select a convenient fraction of 100 and base subsequent steps on this proportion.) Next, count the number of complete sentences and individual syllables in the text sample. These values are used as entry points for a graph that provides grade equivalents for different combinations of sentence and syllable counts from the 100-word sample. The graph indicates that the average number of words and syllables per sentence for a fifth-grade reading level are 14 and 18, respectively. An average sentence at the sixth-grade level has 15 or 16 words and a total of 20 syllables; a seventh-grade-level sentence has about 18 words and 24 syllables. Shorter sentences with a higher proportion of longer words or longer sentences with fewer long words can yield an equivalent grade level. For example, a sentence of 9 words and 13 syllables (i.e., as many as 44% polysyllabic words) and one with 19 words and 22 syllables (i.e., no more than about 14% polysyllabic words) are both classified as sixth-grade reading level. Aiming for a reading level between the fifth and seventh grades is probably an appropriate target for most instruments that will be used with the general population. The items of the Multidimensional Health Locus of

Control (MHLC) scales, for example, were written at a fifth- to seventh-grade reading level. A typical item at this reading level is: "Most things that affect my health happen to me by accident" (Wallston et al., 1978). Its 11 words and 15 syllables place it at the sixth-grade level.

Fry (1977) notes that semantic and syntactic factors should be considered in assessing reading difficulty. Because short words tend to be more common and short sentences tend to be syntactically simpler, his procedure is an acceptable alternative to more complex difficulty assessment methods. However, as with other criteria for writing or choosing good items, one must use common sense in applying reading level methods. Some brief phrases containing only short words are not elementary. "Eschew casque scorn," for example, is more likely to confuse someone with a grade-school education than "Wear your helmet" will, despite the fact that both have 3 words and 4 syllables. Another source of potential confusion that should be avoided is *multiple negatives*: "I am not in favor of corporations stopping funding for antinuclear groups" is much more confusing than "I favor continued private support of groups advocating a nuclear ban." (It is also instructive to observe that these two statements might convey different positions on the issue. For example, the latter might imply a preference for private over public support of the groups in question.)

What are called *double barreled* items should also be avoided. These are items that convey two or more ideas, so that an endorsement of the item might refer to either or both ideas. "I support civil rights because discrimination is a crime against God" is an example of a double-barreled item. If a person supports civil rights for reasons other than its affront to a deity (e.g., because it is a crime against humanity), how should he or she answer? A negative answer might incorrectly convey a lack of support for civil rights, and a positive answer might incorrectly ascribe a motive to the respondent's support.

Another problem that scale developers should avoid is *ambiguous pronoun references*. "Murderers and rapists should not seek pardons from politicians because they are the scum of the earth" might express the sentiments of some people, irrespective of pronoun reference. (However, a scale developer usually intends to be more clear about what an item means.) This sentence should be twice cursed. In addition to the ambiguous pronoun reference, it is double-barreled. *Misplaced modifiers* create ambiguities similar to ambiguous pronoun references: "Our representatives should work diligently to legalize prostitution in the House of Representatives" is an example of such modifiers. Using *adjective forms instead of noun forms* can also create unintended confusion. Consider the differences in meaning between "All vagrants should be given a schizophrenic assessment" and "All vagrants should be given a schizophrenia assessment."

Individual words are not the only sources of item ambiguity. Entire sentences can have more than one meaning. I have actually seen one survey of adolescent sexual behavior that included an item to assess parental education. Given the context of the survey as a whole, the wording was unfortunate: "How far did your mother go in school?" The investigators had totally failed to recognize the unintended meaning of this statement until it evoked snickers from a group of professionals during a seminar presentation. I suspect that a fair number of the adolescent respondents also got a laugh from the item. How it affected the adolescents' responses to the remainder of the questionnaire is unknown.

Positively and Negatively Worded Items

Many scale developers choose to write *negatively worded items,* items that represent low levels or even the absence of the construct of interest, as well as the more common *positively worded items* that represent its presence. The goal is to arrive at a set of items that includes some items that indicate a high level of the latent variable when endorsed and others that indicate a high level when not endorsed. The Rosenberg Self-Esteem (RSE) scale (Rosenberg, 1965), for example, includes items indicative of high esteem (e.g., "I feel that I have a number of good qualities") and of low esteem (e.g., "I certainly feel useless at times"). The intent of wording items both positively and negatively within the same scale is usually to avoid an *acquiescence, affirmation, or agreement bias.* These interchangeable terms refer to a respondent's tendency to agree with items, irrespective of their content. If, for example, a scale consists of items that express a high degree of self-esteem, then an acquiescence bias would result in a pattern of responses appearing to indicate very high esteem. If the scale is made up of equal numbers of positively and negatively worded items, on the other hand, then an acquiescence bias and an extreme degree of self-esteem could be differentiated from one another by the pattern of responses. An "agreer" would endorse items indicating both high and low self-esteem, whereas a person who truly had high esteem would strongly endorse high-esteem items and negatively endorse low-esteem items.

Unfortunately, there may be a price to pay for including positively and negatively worded items. Reversals in item polarity may be confusing to respondents, especially when they are completing a long questionnaire. In such a case, the respondents may become confused about the difference between expressing their strength of agreement with a statement, regardless of its polarity, and expressing the strength of the attribute being measured (esteem, for example). As an applied social science researcher, I have seen many examples of items worded in the opposite direction performing poorly. For example, in DeVellis and Callahan (1993), my colleague and I described

a shorter, more focused alternative to the Rheumatology Attitudes Index (an unfortunate name, as the instrument does not assess attitudes and is not an index). We selected items from the original, longer version based on empirical criteria and ended up with four items expressing negative reactions to illness and one expressing the ability to cope well with illness. The intent was that users should reverse score the "coping" item so that all items expressed a sense of helplessness. More recently, when Currey, Callahan, and DeVellis (2002) examined the performance of that single item worded in the positive direction, the item was found to consistently perform poorly. When the item was reworded by adding the single word "not" to change its valence so as to be consistent with other items, its performance improved dramatically.

We suspect that, although many respondents recognized the different valence of the original item, others did not. This would result in a portion of individuals for whom the original item had positive correlations with the other four items and another portion for whom the same correlations were negative. As a consequence, for the sample as a whole, correlations of that item with the other four would be markedly diminished and thus produce the type of unsatisfactory performance we observed for the original, opposite-valence, item. Personal experience with community-based samples suggests to me that the disadvantages of items worded in an opposite direction outweigh any benefits.

Conclusion

An item pool should be a rich source from which a scale can emerge. It should contain a large number of items that are relevant to the content of interest. Redundancy with respect to content is an asset, not a liability. It is the foundation of internal consistency reliability that, in turn, is the foundation of validity. Items should not involve a "package deal" that makes it impossible for respondents to endorse one part of the item without endorsing another part that may not be consistent with the first. Whether or not positively and negatively worded items are both included in the pool, their wording should follow established rules of grammar. This will help to avoid some of the sources of ambiguity discussed above.

STEP 3: DETERMINE THE FORMAT FOR MEASUREMENT

Numerous formats for questions exist. The researcher should consider early on what the format will be. This step should occur simultaneously with the

generation of items so that the two are compatible. For example, generating a long list of declarative statements may be a waste of time if the response format eventually chosen is a checklist comprised of single-word items. Furthermore, the theoretical models presented earlier are more consistent with some response formats than with others. In general, scales made up of items that are scorable on some continuum and that are summed to form a scale score are most compatible with the theoretical orientation presented in this volume. In this section, however, I will discuss common formats that depart from the pattern implied by the theoretical models discussed in Chapter 2, as well as ones that adhere to that pattern.

Thurstone Scaling

There are a number of general strategies for constructing scales that influence the format of items and response options. One method is *Thurstone scaling*. An analogy may help to clarify how Thurstone scaling works. A tuning fork is designed to vibrate at a specific frequency. If you strike it, it will vibrate at that frequency and produce a specific tone. Conversely, if you place the fork near a tone source that produces the same frequency as the tuning fork, the fork will begin to vibrate. In a sense, then, a tuning fork is a "frequency detector," vibrating in the presence of sound waves of its resonant frequency and remaining motionless in the presence of all other frequencies. Imagine a series of tuning forks aligned in an array such that as one moves from left to right along the array, the tuning forks correspond to progressively higher frequency sounds. Within the range of the tuning forks' frequency, this array can be used to identify the frequency of a tone. In other words, you could identify the tone's frequency by seeing which fork vibrated when the tone was played. A Thurstone scale is intended to work in the same way. The scale developer attempts to generate items that are differentially responsive to specific levels of the attribute in question. When the "pitch" of a particular item matches the level of the attribute a respondent possesses, the item will signal this correspondence. Often, the "signal" consists of an affirmative response for items that are "tuned" to the appropriate level of the attribute and a negative response for all other items. The "tuning" (i.e., determination of to what level of the construct each item responds) is typically determined by having judges place a large pool of items into piles corresponding to equally spaced intervals of construct magnitude or strength.

This is quite an elegant idea. Items could be developed that correspond to different intensities of the attribute, could be spaced to represent equal intervals, and could be formatted with agree-disagree response options, for example. The investigator could give these items to respondents and then inspect

their responses to see which items triggered agreement. Because the items would have been precalibrated with respect to their sensitivity to specific levels of the phenomenon, the agreements would pinpoint how much of the attribute the respondent possessed. The selection of items to represent equal intervals across items would result in highly desirable measurement properties because scores would be amenable to mathematical procedures based on interval scaling.

Part of a hypothetical Thurstone scale for measuring parents' aspirations for their children's educational and career attainments might look like the following:

1. Achieving success is the only way for my child to repay my efforts as a parent	*Agree*_____	*Disagree*_____
2. Going to a good college and getting a good job are important but not essential to my child's happiness	*Agree*_____	*Disagree*_____
3. Happiness has nothing to do with achieving educational or material goals	*Agree*_____	*Disagree*_____
4. The customarily valued trappings of success are a hindrance to true happiness	*Agree*_____	*Disagree*_____

As Nunnally (1978) points out, developing a true Thurstone scale is considerably harder than describing one. Finding items that consistently "resonate" to specific levels of the phenomenon is quite difficult. The practical problems associated with the method often outweigh its advantages—unless the researcher has a compelling reason for wanting the type of calibration that it provides. Although Thurstone scaling is an interesting and sometimes suitable approach, it will not be referred to herein henceforth. Note, however, that methods based on item response theory, discussed in Chapter 7, share many of the goals of Thurstone scales while taking a somewhat different approach to achieving them.

Guttman Scaling

A *Guttman scale* is a series of items tapping progressively higher levels of an attribute. Thus a respondent should endorse a block of adjacent items until, at a critical point, the amount of the attribute that the items tap exceeds that possessed by the subject. None of the remaining items should be endorsed.

Some purely descriptive data conform to a Guttman scale. For example, a series of interview questions might ask, "Do you smoke?" "Do you smoke more than 10 cigarettes a day?" "Do you smoke more than a pack a day?" and so on. As with this example, endorsing any specific item on a Guttman scale implies affirmation of all preceding items. A respondent's level of the attribute is indicated by the highest item yielding an affirmative response. Note that, whereas both Thurstone and Guttman scales are made up of graded items, the focus is on a single affirmative response in the former case, but the point of transition from affirmative to negative responses is the focus of the latter. A Guttman version of the preceding parental aspiration scale might look like this:

1. Achieving success is the only way for my child to repay my efforts as a parent	*Agree*_____	*Disagree*_____
2. Going to a good college and getting a good job are very important to my child's happiness	*Agree*_____	*Disagree*_____
3. Happiness is more likely if a person has attained his or her educational and material goals	*Agree*_____	*Disagree*_____
4. The customarily valued trappings of success are not a hindrance to true happiness	*Agree*_____	*Disagree*_____

Guttman scales can work quite well for objective information or in situations where it is a logical necessity that responding positively to one level of a hierarchy implies satisfying the criteria of all lower levels of that hierarchy. Things get murkier when the phenomenon of interest is not concrete. In the case of our hypothetical parental aspiration scale, for example, the ordering may not be uniform across individuals. Whereas 20 cigarettes a day always implies more smoking than 10, responses to items 3 and 4 in the parental aspiration scale example may not always conform to the ordering pattern of a Guttman scale. For example, a person might agree with item 3 but disagree with item 4. Ordinarily, agreement with item 3 would imply agreement with 4, but if a respondent viewed success as a complex factor that acted simultaneously as a help and a hindrance to happiness, then an atypical pattern of responses could result.

Like Thurstone scales, Guttman scales undoubtedly have their place, but their applicability seems rather limited. With both approaches, the disadvantages and difficulties will often outweigh the advantages. It is important to

remember that the measurement theories discussed thus far do not always apply to these types of scales. Certainly, the assumption of equally strong causal relationships between the latent variable and each of the items would not apply to Thurstone or Guttman scale items. Nunnally and Bernstein (1994) describe briefly some of the conceptual models underlying these scales. For situations in which ordered items are particularly appropriate, IRT-based models, discussed in Chapter 7, are potentially an appropriate choice, although implementing these methods can be quite burdensome.

Scales With Equally Weighted Items

The measurement models discussed earlier fit best with scales consisting of items that are more or less equivalent "detectors" of the phenomenon of interest—that is, they are more or less parallel (but not necessarily parallel in the strict sense of the parallel tests model). They are imperfect indicators of a common phenomenon that can be combined by simple summation into an acceptably reliable scale.

One attractive feature of scales of this type is that the individual items can have a variety of response option formats. This allows the scale developer a good deal of latitude in constructing a measure optimally suited for a particular purpose. Some general issues related to response formatting will be examined below, as will the merits and liabilities of some representative response formats.

What is the Optimum Number of Response Categories?

Most scale items consist of two parts: a stem and a series of response options. For example, the stem of each item may be a different declarative statement expressing an opinion, and the response options accompanying each stem might be a series of descriptors indicating the strength of agreement with the statement. For now, let us focus on the response options—specifically, the number of choices that should be available to the respondent. Some item response formats allow the subject an infinite or very large number of options, whereas others limit the possible responses. Imagine, for example, a response scale for measuring anger that resembles a thermometer, calibrated from "no anger at all" at the base of the thermometer to "complete, uncontrollable rage" at its top. A respondent could be presented with a series of situation descriptions, each accompanied by a copy of the thermometer scale, and asked to indicate, by shading in some portion of the thermometer, how much anger the situation provoked. This method allows for virtually continuous measurement

of anger. An alternative method might ask the respondent to indicate, using a number from 1 to 100, how much anger each situation provoked. This provides for numerous discrete responses. Alternatively, the format could restrict the response options to a few choices, such as "none," "a little," "a moderate amount," and "a lot," or to a simple binary selection between "angry" and "not angry."

What are the relative advantages of these alternatives? A desirable quality of a measurement scale is variability. A measure cannot covary if it does not vary. If a scale fails to discriminate differences in the underlying attribute, its correlations with other measures will be restricted and its utility will be limited. One way to increase opportunities for variability is to have lots of scale items. Another is to have numerous response options within items. If circumstances restrict an investigator to two questions regarding anger, for example, it might be best to allow respondents latitude in describing their level of anger. Assume that the research concerns the enforcement of nonsmoking policies in a work setting. Let us further assume that the investigators want to determine the relationship between policy and anger. If the investigators were limited to only two questions (e.g., "How much anger do you feel when you are restricted from smoking?" and "How much anger do you feel when you are exposed to others smoking in the work place?"), they might get more useful information from a response format that allowed subjects many gradations of response than from a binary response format. For example, a 0 to 100 scale might reveal wide differences in reactions to these situations and yield good variability for the two-item scale. On the other hand, if the research team were allowed to include 50 questions about smoking and anger, simple "angry" versus "not angry" indications might yield sufficient variability when the items are added to obtain a scale score. In fact, being faced with more response options on each of 50 questions might fatigue or bore the respondents, lowering the reliability of their responses.

Another issue related to the number of response options is the *respondents' ability to discriminate meaningfully.* How fine a distinction can the typical subject make? This obviously depends on what is being measured. Few things can truly be evaluated into, say, 50 discrete categories. Presented with this many options, many respondents may use only those corresponding to multiples of 5 or 10, effectively reducing the number of options to as few as five. Differences between a response of 35 and 37 may not reflect actual difference in the phenomenon being measured. Little is gained with this sort of false precision. Although the scale's variance might increase, it may be the random (i.e., error) portion rather than the systematic portion attributable to the underlying phenomenon that is increasing. This, of course, offers no benefit.

Sometimes the respondent's ability to discriminate meaningfully between response options will depend on the *specific wording or physical placement* of those options. Asking a respondent to discriminate among vague quantity descriptors, such as "several," "few," and "many," may create problems. Sometimes the ambiguity can be reduced by the arrangement of the response options on the page. Respondents often seem to understand what is desired when they are presented with an obvious continuum. Thus an ordering such as

Many *Some* *Few* *Very Few* *None*

may imply that "some" is more than "few" because of the ordering of these items. However, if it is possible to find a nonambiguous adjective that precludes the respondents' making assumptions based on location along a continuum, so much the better. At times, it may be preferable to have fewer response options than to have ones that are ambiguous. So, for example, it may be better in the above example to eliminate either "some" or "few" and have four options rather than five. The worst circumstance is to combine ambiguous words with ambiguous page locations. Consider the following example:

Very Helpful *Not Very Helpful*

Somewhat Helpful *Not at all Helpful*

Terms such as "somewhat" and "not very" are difficult to differentiate under the best of circumstances. However, arranging these response options as they appear above makes matters even worse. If a respondent reads down the first column and then down the second, "somewhat" appears to represent a higher value than "not very." But, if a respondent reads across the first row and then across the second, the implicit ordering of these two descriptors along the continuum is reversed. Due to ambiguity in both language and spatial arrangement, individuals may assign different meanings to the two options representing moderate values, and reliability would suffer as a consequence.

Still another issue is the *investigator's ability and willingness to record a large number of values* for each item. If the thermometer method described earlier is used to quantify anger responses, is the researcher actually going to attempt a precise scoring of each response? How much precision is appropriate? Can the shaded area be measured to within a quarter of an inch? A centimeter? A millimeter? If only some crude datum, say lower, middle, or upper third, is extracted from the scale, what was the point in requesting such a precise response?

There is at least one more issue related to the number of responses. Assuming that a few discrete responses are allowed for each item, *should the number be odd or even?* Again, this depends on the type of question, the type of response option, and the investigator's purpose. If the response options are bipolar, with one extreme indicating the opposite of the other (e.g., a strong positive versus a strong negative attitude), an odd number of responses permits equivocation (e.g., "neither agree nor disagree") or uncertainty (e.g., "not sure"); an even number usually does not. An odd number implies a central "neutral" point (e.g., neither a positive nor a negative appraisal). An even number of responses, on the other hand, forces the respondent to make at least a weak commitment in the direction of one or the other extreme (e.g., a forced choice between a mildly positive or mildly negative appraisal as the least extreme response). Neither format is necessarily superior. The researcher may want to preclude equivocation if it is felt that subjects will select a neutral response as a means of avoiding a choice. In studies of social comparison choices, for example, the investigators may want to force subjects to express a preference for information about a more advantaged or less advantaged person. Consider these two alternative formats, the first of which was chosen for a study of social comparisons among people with arthritis (DeVellis et al., 1990):

Would you prefer information about:

(a) Patients who have worse arthritis than you have
(b) Patients who have milder arthritis than you have

Would you prefer information about:

(a) Patients who have worse arthritis than you have
(b) Patients who have arthritis equally as bad as you have
(c) Patients who have milder arthritis than you have

A neutral option such as 2b might permit unwanted equivocation. A neutral point may also be desirable. In a study assessing which of two risks (e.g., boredom or danger) people prefer taking, a midpoint may be crucial. The researcher might vary the chance or severity of harm across several choices between a safe, dull activity and an exciting, risky one. The point at which a respondent is most nearly equivocal about risking the more exciting activity could then be used as an index of risk-taking:

Indicate your relative preference for activity A or activity B from the alternatives listed below by circling the appropriate phrase following the description of activity B.

Activity A: Reading a statistics book (no chance of severe injury)

1. Activity B: Taking a flight in a small commuter plane (very slight chance of severe injury)

Strongly Prefer A	*Mildly Prefer A*	*No Preference*	*Mildly Prefer B*	*Strongly Prefer B*

2. Activity B: Taking a flight in a small open-cockpit plane (slight chance of severe injury)

Strongly Prefer A	*Mildly Prefer A*	*No Preference*	*Mildly Prefer B*	*Strongly Prefer B*

3. Activity B: Parachute jumping from a plane with a backup chute (moderate chance of severe injury)

Strongly Prefer A	*Mildly Prefer A*	*No Preference*	*Mildly Prefer B*	*Strongly Prefer B*

4. Activity B: Parachute jumping from a plane without a backup chute (substantial risk of severe injury)

Strongly Prefer A	*Mildly Prefer A*	*No Preference*	*Mildly Prefer B*	*Strongly Prefer B*

5. Activity B: Jumping from a plane without a parachute and attempting to land on a soft target (almost certain severe injury)

Strongly Prefer A	*Mildly Prefer A*	*No Preference*	*Mildly Prefer B*	*Strongly Prefer B*

The other merits or liabilities of this approach aside, it would clearly require that response options include a midpoint.

Specific Types of Response Formats

Scale items occur in a dizzying variety of forms. However, there are several ways to present items that are used widely and have proven successful in diverse applications. Some of these are discussed below.

Likert Scale

One of the most common item formats is a *Likert scale.* When a Likert scale is used, the item is presented as a declarative sentence, followed by response

options that indicate varying degrees of agreement with or endorsement of the statement. (In fact, the preceding example of risk-taking used a Likert response format.) Depending on the phenomenon being investigated and the goals of the investigator, either an odd or an even number of response options might accompany each statement. The response options should be worded so as to have roughly equal intervals with respect to agreement. That is to say, the difference in agreement between any adjacent pair of responses should be about the same as it is for any other adjacent pair of response options. A common practice is to include six possible responses: "strongly disagree," "moderately disagree," "mildly disagree," "mildly agree," "moderately agree," and "strongly agree." These form a continuum from strong disagreement to strong agreement. A neutral midpoint can also be added. Common choices for a midpoint include "neither agree nor disagree" and "agree and disagree equally." There is legitimate room for discussion concerning the equivalence of these two midpoints. The first implies apathetic disinterest, while the latter suggests strong but equal attraction to both agreement and disagreement. It may very well be that most respondents do not focus very much attention on subtleties of language but merely regard any reasonable response option in the center of the range as a midpoint irrespective of its precise wording.

Likert scaling is widely used in instruments measuring opinions, beliefs, and attitudes. It is often useful for these statements to be fairly (though not extremely) strong when used in a Likert format. Presumably, the moderation of opinion is expressed in the choice of response option. For example, the statements "Physicians generally ignore what patients say," "Sometimes, physicians do not pay as much attention as they should to patients' comments," and "Once in a while, physicians might forget or miss something a patient has told them" express strong, moderate, and weak opinions, respectively, concerning physicians' inattention to patients' remarks. Which is best for a Likert scale? Ultimately, of course, the one that most accurately reflects true differences of opinion is best. In choosing how strongly to word items in an initial item pool, the investigator might profitably ask, "How are people with different amounts or strengths of the attribute in question likely to respond?" In the case of the three examples just presented, the investigator might conclude that the last question would probably elicit strong agreement from people whose opinions fell along much of the continuum from positive to negative. If this conclusion proved correct, then the third statement would not do a good job of differentiating between people with strong and moderate negative opinions.

In general, very mild statements may elicit too much agreement when used in Likert scales. Many people will strongly agree with such a statement as "The safety and security of citizens are important." One could strongly agree

with such a statement (i.e., choose an extreme response option) without holding an extreme opinion. Of course, the opposite is equally true. People holding any but the most extreme views might find themselves in disagreement with an extremely strong statement (for example, "Hunting down and punishing wrongdoers is more important than protecting the rights of individuals"). Of the two (overly mild or overly extreme) statements, the former may be the bigger problem for two reasons. First, our inclination is often to write statements that will not offend our subjects. Avoiding offensiveness is probably a good idea. However, it may lead us to favor items that nearly everyone will find agreeable. Another reason to be wary of items that are too mild is that they may represent the absence of belief or opinion. The third of our inattentive physician items in the preceding paragraph did not indicate the presence of a favorable attitude so much as the absence of an unfavorable one. Items of this sort may be poorly suited to the research goal because we are more often interested in the presence of some phenomenon rather than in its absence.

In summary, a good Likert item should state the opinion, attitude, belief, or other construct under study in clear terms. It is neither necessary nor appropriate for this type of scale to span the range of weak to strong assertions of the construct. The response options provide the opportunity for gradations.

An example of items in Likert response formats is as follows:

1. Exercise is an essential component of a healthy lifestyle.

1	2	3	4	5	6
Strongly Disagree	*Moderately Disagree*	*Mildly Disagree*	*Mildly Agree*	*Moderately Agree*	*Strongly Agree*

2. Combating drug abuse should be a top national priority.

1	2	3	4	5
Completely True	*Mostly True*	*Equally True and Untrue*	*Mostly Untrue*	*Completely Untrue*

Semantic Differential

The semantic differential scaling method is chiefly associated with the attitude research of Osgood and his colleagues (e.g., Osgood & Tannenbaum, 1955). Typically, a *semantic differential* is used in reference to one or more stimuli. In the case of attitudes, for example, the stimulus might be a group of people, such as automobile salespeople. Identification of the target stimulus is followed by a list of adjective pairs. Each pair represents opposite ends of a

continuum, defined by adjectives (e.g., honest and dishonest). As shown in the example below, there are several lines between the adjectives that constitute the response options.

Automobile Salesperson

Honest	____	____	____	____	____	____	____	*Dishonest*
Quiet	____	____	____	____	____	____	____	*Noisy*

In essence, the individual lines (seven and nine are common numbers) represent points along the continuum defined by the adjectives. The respondent places a mark on one of the lines to indicate the point along the continuum that characterizes the stimulus. For example, if someone regarded auto salespeople as extremely dishonest, he or she might select the line closest to that adjective. Either extreme or moderate views can be expressed by choosing which line to mark. After rating the stimulus with regard to the first adjective pair, the person would proceed to additional adjective pairs separated by lines.

The adjectives one chooses can be either bipolar or unipolar, depending, as always, on the logic of the research questions the scale is intended to address. Bipolar adjectives each express the presence of opposite attributes, such as friendly and hostile. Unipolar adjective pairs indicate the presence and absence of a single attribute, such as friendly and not friendly.

Like the Likert scale, the semantic differential response format can be highly compatible with the theoretical models presented in the earlier chapters of this book. Sets of items can be written to tap the same underlying variable. For example, items using trustworthy/untrustworthy, fair/unfair, and truthful/untruthful as endpoints might be added to the first statement in the preceding example to constitute an "honesty" scale. Such a scale could be conceptualized as a set of items sharing a common latent variable (honesty) and conforming to the assumptions discussed in Chapter 2. Accordingly, the scores of the individual "honesty" items could be added and analyzed as described in a later section concerning the evaluation of items.

Visual Analog

Another item format that is in some ways similar to the semantic differential is the *visual analog scale.* This response format presents the respondent with a continuous line between a pair of descriptors representing opposite ends of a continuum. The individual completing the item is instructed to place a mark at a point on the line that represents his or her opinion, experience, belief, or whatever is being measured. The visual analog scale, as the term *analog* in the name implies, is a continuous scale. The fineness of differentiation in assigning scores to

points on the scale is determined by the investigator. Some of the advantages and disadvantages of a continuous response format were discussed earlier. An additional issue not raised at that time concerns possible differences in the interpretation of physical space as it relates to values on the continuum. A mark placed at a specific point along the line may not mean the same thing to different people, even when the end points of the line are identically labeled for all respondents. Consider a visual analog scale for pain such as this:

No Pain At All _____ *Worst Pain I Ever Experienced*

Does a response in the middle of the scale indicate pain about half of the time, constant pain of half the possible intensity, or something else entirely? Part of the problem with measuring pain is that pain can be evaluated on multiple dimensions, including frequency, intensity, and duration. Also, recollections of the worst pain a given person has ever experienced are likely to be distorted. Comparisons across individuals are further complicated by the fact that different people may have experienced different levels of "the worst pain." Of course, some of these problems reside with the phenomenon used in this example, pain (see Keefe, 2000, for an excellent discussion of pain measurement), and not with the scale per se. However, the problem of idiosyncratic assignment of values along a visual analog scale can exist for other phenomena as well.

A major advantage of visual analog scales is that they are potentially very sensitive (Mayer, 1978). This can make them especially useful for measuring phenomena before and after some intervening event, such as an intervention or experimental manipulation, that exerts a relatively weak effect. A mild rebuke in the course of an experimental manipulation, for example, may not produce a shift on a 5-point measure of self-esteem. However, a subtle but systematic shift to lower values on a visual analog scale might occur among people in the "rebuke" condition of this hypothetical experiment. Sensitivity may be more advantageous when examining changes over time within the same individual rather than changes across individuals (Mayer, 1978). This may be so because, in the former case, there is no error added due to extraneous differences between individuals.

Another potential advantage of visual analog scales when they are repeated over time is that it is difficult or impossible for subjects to encode their past responses with precision. To continue with the example from the preceding paragraph, a subject would probably have little difficulty remembering which of five numbered options to a self-esteem item he or she had previously chosen in response to a multiresponse format such as a Likert scale. Unless

one of the end points of a visual analog scale were chosen, however, it would be very difficult to recall precisely where a mark had been made along a featureless line. This could be advantageous if the investigator were concerned that respondents might be biased to appear consistent over time. Presumably, subjects motivated to be consistent would choose the same response after exposure to an experimental intervention as prior to such exposure. The visual analog format essentially rules out this possibility. If the postmanipulation responses departed consistently (i.e., usually in the same direction) from the premanipulation response for experimental subjects and randomly for controls, then the choice of a visual analog scale might have contributed to detecting a subtle phenomenon that other methods would have missed.

Visual analog scales have often been used as single-item measures. This has the sizeable disadvantage of precluding any determination of internal consistency. With a single-item measure, reliability can be determined only by the test-retest method described in Chapter 3 or by comparison with other measures of the same attribute having established psychometric properties. The former method suffers from the problems of test-retest assessments discussed earlier, notably the impossibility of differentiating instability of the measurement process from instability of the phenomenon being measured. The latter method is actually a construct validity comparison. However, because reliability is a necessary condition for validity, one can infer the reliability if validity is in evidence. Nonetheless, a better strategy may be to develop multiple visual analog items so that internal consistency can be determined.

Numerical Response Formats and Basic Neural Processes

A recent study by Zorzi, Priftis, and Umilitá (2002) that appeared in *Nature* suggests that certain response options may correspond to how the brain processes numerical information. According to these authors, numbers arrayed in a sequence, as with the typical Likert scale, express quantity not only in their numerical values but in their locations. They suggest that the visual line of numbers is not merely a convenient representation but corresponds to fundamental neural processes. They observed that people with various brain lesions that impair spatial perception in the visual field make systematic errors in simple, visually presented mathematical problems. The spatial anomaly and the type of errors are very closely linked. Individuals unable to perceive the left visual field who were asked to indicate the midpoint between two values presented in a linear array consistently erred "to the right." For example, when asked what would be midway between points labeled "3" and "9", there were errors shifted to the right (i.e., to higher values). Reversing the scale from high to low continued to produce shifts to the right (now, lower values). When the

same tasks were presented in nonvisual form—for example, by asking what the average of 3 and 9 was—the pattern did not appear. In fact, these individuals showed no deficit in performing arithmetic when it was not presented visually. Control subjects without the visual anomaly did not show the shift pattern of those with brain lesions. The authors conclude that their work constitutes, "strong evidence that the mental number line is more than simply a metaphor" and that "thinking of numbers in spatial terms (as has been reported by great mathematicians) may be more efficient because it is grounded in the actual neural representation of numbers" (Zorzi et al., 2002, p. 138). Although this study, by itself, may not warrant hard and fast conclusions, it provides tantalizing preliminary evidence that evaluating a linear string of numbers may correspond to fundamental neural mechanisms involved in assessing quantity. If this is truly the case, then response options presented as a row of numbers may have special merit.

Binary Options

Another common response format gives subjects a choice between *binary options* for each item. The earlier examples of Thurstone and Guttman scales used binary options ("agree" and "disagree"), although scales with equally weighted items could also have binary response options. Subjects might, for example, be asked to check off all the adjectives on a list that they think apply to themselves. Or, they may be asked to answer "yes" or "no" to a list of emotional reactions they may have experienced in some specified situation. In both cases, responses reflecting items sharing a common latent variable (e.g., adjectives such as "sad," "unhappy," and "blue" that represent depression) could be combined into a single score for that construct.

A major shortcoming of binary responses is that each item can have only minimal variability. Similarly, any pair of items can have only one of two levels of covariation: agreement or disagreement. Recall from Chapter 3 that the variance of a scale made up of multiple equally weighted items is exactly equal to the sum of all the elements in the covariance matrix for the individual items. With binary items, each item contributes precious little to that sum because of the limitations in possible variances and covariances. The practical consequence of this is that more items are needed to obtain the same degree of scale variance if the items are binary. However, binary items are usually extremely easy to answer. Therefore, the burden placed on the subject is very low for any one item. For example, most people can quickly decide whether certain adjectives are apt descriptions of themselves. As a result, subjects often are willing to complete more binary items than ones using a format demanding concentration on finer distinctions. Thus a binary format may allow the investigator to achieve adequate variation in scale scores by aggregating information over more items.

Item Time Frames

Another issue that pertains to the formatting of items is the time frame specified or implied. Kelly and McGrath (1988), in another volume in this series, have discussed the importance of considering the temporal features of different measures. Some scales will not make reference to a time frame, implying a universal time perspective. Locus of control scales, for example, often contain items that imply an enduring belief about causality. Items such as "If I take the right actions, I can stay healthy" (Wallston et al., 1978) presume that this belief is relatively stable. This is consistent with the theoretical characterization of locus of control as a generalized rather than specific expectancy for control over outcomes (although there has been a shift toward greater specificity in later measures of locus of control beliefs—e.g., DeVellis, DeVellis, Revicki, Lurie, Runyan, & Bristol, 1985). Other measures assess relatively transient phenomena. Depression, for example, can vary over time and scales to measure it have acknowledged this point (Mayer, 1978). For example, the widely used Center for Epidemiological Studies Depression (CES-D) scale (Radloff, 1977) uses a format that asks respondents to indicate how often during the past week they have experienced various mood states. Some measures, such as anxiety scales (e.g., Spielberger, Gorsuch, & Lushene, 1970), are developed in different forms intended to assess relatively transient states or relatively enduring traits (Zuckerman, 1983). The investigator should choose a time frame for a scale actively rather than passively. Theory is an important guide to this process. Is the phenomenon of interest a fundamental and enduring aspect of individuals' personalities, or is it likely to be dependent on changing circumstances? Is the scale intended to detect subtle variations occurring over a brief time frame (e.g., increases in negative affect after viewing a sad movie) or changes that may evolve over a lifetime (e.g., progressive political conservatism with increasing age)?

In conclusion, the item formats, including response options and instructions, should reflect the nature of the latent variable of interest and the intended uses of the scale.

STEP 4: HAVE THE INITIAL ITEM POOL REVIEWED BY EXPERTS

Thus far, we have examined the need for clearly articulating what the phenomenon of interest is, generating a pool of suitable items, and selecting a response

format for those items. The next step in the process is asking a group of people who are knowledgeable in the content area to review the item pool. This review serves multiple purposes related to maximizing the content validity (see Chapter 4) of the scale.

First, having experts review your item pool can confirm or invalidate your definition of the phenomenon. You can ask your panel of experts (e.g., colleagues who have worked extensively with the construct in question or related phenomena) to rate *how relevant they think each item is to what you intend to measure.* This is especially useful if you are developing a measure that will consist of separate scales to measure multiple constructs. If you have been careful in developing your items, then experts should have little trouble determining which items correspond to which constructs. In essence, your thoughts about what each item measures is the hypothesis, and the responses of the experts are the confirming or disconfirming data. Even if all the items are intended to tap a single attribute or construct, expert review is useful. If experts read something into an item that you did not plan to include, subjects completing a final scale might do likewise.

The mechanics of obtaining evaluations of item relevance usually involve providing the expert panel with your working definition of the construct. They are then asked to rate each item with respect to its relevance vis-à-vis the construct as you have defined it. This might entail merely rating relevance as high, moderate, or low for each item. In addition, you might invite your experts to comment on individual items as they see fit. This makes their job a bit more difficult but can yield excellent information. A few insightful comments about why certain items are ambiguous, for example, might give you a new perspective on how you have attempted to measure the construct.

Reviewers also can *evaluate the items' clarity and conciseness.* The content of an item may be relevant to the construct, but its wording may be problematic. This bears on item reliability because an ambiguous or otherwise unclear item, to a greater degree than a clear item, can reflect factors extraneous to the latent variable. In your instructions to reviewers, ask them to point out awkward or confusing items and suggest alternative wordings, if they are so inclined.

A third service that your expert reviewers can provide is *pointing out ways of tapping the phenomenon that you have failed to include.* There may be a whole approach that you have overlooked. For example, you may have included many items referring to illness in a pool of items concerned with health beliefs but failed to consider injury as another relevant departure from health. By reviewing the variety of ways you have captured the phenomenon of interest, your reviewers can help you to maximize the content validity of your scale.

A final word of caution concerning expert opinion: The final decision to accept or reject the advice of your experts is your responsibility as the scale developer. Sometimes content experts might not understand the principles of scale construction. This can lead to bad advice. A recommendation I have frequently encountered from colleagues without scale development experience is to eliminate items that concern the same thing. As discussed earlier, removing all redundancy from an item pool or a final scale would be a grave error because redundancy is an integral aspect of internal consistency. However, this comment might indicate that the wording, vocabulary, and sentence structure of the items are too similar and could be improved. Pay careful attention to all the suggestions you receive from content experts. Then make your own informed decisions about how to use their advice.

At this point in the process, the scale developer has a set of items that has been reviewed by experts and modified accordingly. It is now time to advance to the next step.

STEP 5: CONSIDER INCLUSION OF VALIDATION ITEMS

Obviously, the heart of the scale development questionnaire is the set of items from which the scale under development will emerge. However, some foresight can pay off handsomely. It might be possible and relatively convenient to include some additional items in the same questionnaire that will help in determining the validity of the final scale. There are at least two types of items to consider.

The first type of item a scale developer might choose to include in a questionnaire serves to detect flaws or problems. Respondents might not be answering the items of primary interest for the reasons you assume. There may be other motivations influencing their responses. Learning this early is advantageous. One type of motivation that can be assessed fairly easily is *social desirability.* If an individual is strongly motivated to present herself or himself in a way that society regards as positive, item responses may be distorted. Including a social desirability scale allows the investigator to assess how strongly individual items are influenced by social desirability. Items that correlate substantially with the social desirability score obtained should be considered as candidates for exclusion unless there is a sound theoretical reason that indicates otherwise. A brief and useful social desirability scale has been developed by Strahan and Gerbasi (1972). This 10-item measure can be conveniently inserted into a questionnaire.

There are other sources of items for detecting undesirable response tendencies (Anastasi, 1968). The Minnesota Multiphasic Personality Inventory, or MMPI (Hathaway & Meehl, 1951; Hathaway & McKinley, 1967), includes several scales aimed at detecting various response biases. In some instances, it may be appropriate to include these types of scales.

The other class of items to consider including at this stage pertains to the construct validity of the scale. As discussed in Chapter 4, if theory asserts that the phenomenon you are setting out to measure relates to other constructs, then the performance of the scale vis-à-vis measures of those other constructs can serve as evidence of its validity. Rather than mounting a separate validation effort after constituting the final scale, it may be possible to include measures of relevant constructs at this stage. The resultant pattern of relationships can provide support for claims of validity or, alternatively, provide clues as to why the set of items does not perform as anticipated.

STEP 6: ADMINISTER ITEMS TO A DEVELOPMENT SAMPLE

After deciding which construct-related and validity items to include in your questionnaire, you must administer them, along with the pool of new items, to some subjects. The sample of subjects should be large. How large is large? It is difficult to find a consensus on this issue. Let us examine the rationale for a large sample. Nunnally (1978) points out that the primary sampling issue in scale development involves the sampling of items from a hypothetical universe (cf. Ghiselli et al., 1981). In order to concentrate on the adequacy of the items, the sample should be sufficiently large to eliminate subject variance as a significant concern. He suggests that 300 people is an adequate number. However, practical experience suggests that scales have been successfully developed with smaller samples. The number of items and the number of scales to be extracted also have a bearing on the sample size issue. If only a single scale is to be extracted from a pool of about 20 items, fewer than 300 subjects might suffice.

There are several risks in using too few subjects. First, the patterns of covariation among the items may not be stable. An item that appears to increase internal consistency may turn out to be a dud when it is used on a separate sample. If items are selected for inclusion (as they very well may be) on the basis of their contribution to alpha, having a small developmental sample can paint an inaccurately rosy picture of internal consistency. When the ratio of subjects to items is relatively low and the sample size is not large, the

correlations among items can be influenced by chance to a fairly substantial degree. When a scale whose items were selected under these conditions is readministered, the chance factors that made certain items look good initially are no longer operative. Consequently, the alpha obtained on occasions other than the initial development study may be lower than expected. Similarly, a potentially good item may be excluded because its correlation with other items was attenuated purely by chance.

A second potential pitfall of small sample size is that the development sample may not represent the population for which the scale is intended. Of course, this can also be the case if the development sample is large, but a small sample is even more likely to exclude certain types of individuals. Thus a scale developer should consider both the size and the composition of the development sample. A careful investigator might choose to address the generalizability of a scale across populations (or some other facet) with a G-study, as discussed in Chapter 3.

Not all types of nonrepresentativeness are identical. There are at least two different ways in which a sample may not be representative of the larger population. The first involves the level of the attribute present in the sample versus in the intended population. For example, a sample might represent a narrower range of the attribute than would be expected of the population. This constriction of range may also be asymmetrical, so that the mean score obtained on the scale for the sample is appreciably higher or lower than one would expect for the population. Opinions regarding the appropriate legal drinking age, for example, might very well differ on a college campus and in a community at large. A mean value of the attribute that is not representative does not necessarily disqualify the sample for purposes of scale development. It may yield inaccurate expectations for scale means while still providing an accurate picture of the internal consistency the scale possesses. For example, a sample of this sort might still lead to correct conclusions about which items are most strongly interrelated.

A more troublesome type of sample nonrepresentativeness involves a sample that is qualitatively rather than quantitatively different from the target population. Specifically, a sample in which the relationships among items or constructs may differ from those in the population is reason for concern. If a sample is quite unusual, items may have a different meaning than for people in general. The patterns of association among items might reflect unusual attributes shared among sample members but rare in the broader community. In other words, the groupings of interrelated items that emerge (e.g., from a factor analysis) may be atypical. Stated a bit more formally, the underlying causal structure relating variables to true scores may be different if a sample is unlike the population in important ways. Consider some rather

obvious examples: If the members of the sample chosen do not understand a key word that recurs among the items and has relevance to the construct, their responses may tell little or nothing about how the scale would perform under different circumstances. The word *sick* means "ill" in the United States but "nauseated" (i.e., sick to one's stomach) in England. Thus a set of questions about illness developed for one group may have a markedly different meaning for another. If the scale concerns a specific health problem not usually associated with nausea (e.g., arthritis), items that use the word *ill* might cluster together because of their distinct meaning if the sample were British. An American sample, on the other hand, would be unlikely to differentiate statements about being ill from other health-related items. Even within the United States, the same word can have different meanings. Among rural Southerners, for example, *bad blood* is sometimes used as a euphemism for venereal disease, whereas in other parts of the country, it means "animosity." If an item discussing "bad blood between relatives" performed differently in a sample of rural Southerners versus other samples, it would hardly be surprising.

The consequences of this second type of sample nonrepresentativeness can severely harm a scale development effort. The underlying structure that emerges—the patterns of covariation among items that are so important to issues of scale reliability—may be a quirk of the sample used in development. If a researcher has reason to believe that the meaning ascribed to items in a development sample may be atypical of the meaning of those items in the larger population, great caution should be used in interpreting the findings obtained from that sample.

STEP 7: EVALUATE THE ITEMS

After an initial pool of items has been developed, scrutinized, and administered to an appropriately large and representative sample, it is time to evaluate the performance of the individual items so that appropriate ones can be identified to constitute the scale. This is, in many ways, the heart of the scale development process. Item evaluation is second perhaps only to item development in its importance.

Initial Examination of Items' Performance

When discussing item development, we examined some of the qualities that are desirable in a scale item. Let us reconsider that issue. The ultimate

quality we seek in an item is a high correlation with the true score of the latent variable. This follows directly from the discussion of reliability in Chapter 3. We cannot directly assess the true score (if we could, we probably would not need a scale) and thus cannot directly compute its correlations with items. However, we can make inferences, based on the formal measurement models that have been discussed thus far. When discussing parallel tests in Chapter 2, I noted that the correlation between any two items equaled the square of the correlation between either item and the true score. This squared value is the reliability of each of the items. So, we can learn about relationships to true scores from correlations among items. The higher the correlations among items, the higher are the individual item reliabilities (i.e., the more intimately they are related to the true score). The more reliable the individual items are, the more reliable will be the scale that they comprise (assuming that they share a common latent variable). So, the first quality we seek in a set of scale items is that they be *highly intercorrelated.* One way to determine how intercorrelated the items are is to inspect the correlation matrix.

Reverse Scoring

If there are items whose correlations with other items are negative, then the appropriateness of *reverse scoring* those items should be considered. Earlier, I suggested that items worded in opposite directions can pose problems. Sometimes, however, we may inadvertently end up with negatively correlated items. This might happen, for example, if we initially anticipated two separate groups of items (e.g., pertaining to happiness and sadness), but decide for some reason that they should be combined into a single group. We could then wind up with statements that relate equally to the new, combined construct (e.g., affect), but some statements may be positive and others negative. "I am happy" and "I am sad" both pertain to affect. However, they are opposites. If we wanted high scores on our scale to measure happiness, then we would have to ascribe a high value to endorsing the "happy" item but a low value to endorsing the "sad" item. That is to say, we would reverse score the sadness item. Sometimes, items are administered in such a way that they are already reversed. For example, subjects might be asked to circle higher numerical values to indicate agreement with a "happy" item and lower values to endorse a "sad" one. One way to do this is by having the verbal descriptors for the response options (e.g., "strongly disagree," "moderately disagree," etc.) always in the same order for all items, but having the numbers associated with them either ascend or descend, depending on the item, as follows:

1. I am sad often.

6	5	4	3	2	1
Strongly Disagree	*Moderately Disagree*	*Mildly Disagree*	*Mildly Agree*	*Moderately Agree*	*Strongly Agree*

2. Much of the time, I am happy.

1	2	3	4	5	6
Strongly Disagree	*Moderately Disagree*	*Mildly Disagree*	*Mildly Agree*	*Moderately Agree*	*Strongly Agree*

This process may confuse the subject. People may ignore the words after realizing that they are the same for all items. However, it is probably preferable to altering the order of the descriptors (e.g., from "strongly disagree" to "strongly agree," from left to right, for some items and the reverse for others). Another option is to have both the verbal descriptions and their corresponding numbers the same for all items but to enter different values for certain items at the time of data coding. Changing scores for certain items at the time of coding is both tedious and potentially error prone. For every subject, every item to be reverse scored must be given the special attention involved in reverse scoring. This creates numerous opportunities for mistakes.

The easiest method to use for reverse scoring is to do it electronically once the data have been entered into a computer. A few computer statements can handle all the reverse scoring for all subjects' data. If the response options have numerical values and the desired transformation is to reverse the order of values, a simple formula can be used. For example, assume that a set of mood items formatted using a Likert scale was scored from 1 to 7, with higher numbers indicating agreement. Assume further that, for ease of comprehension, both positive mood items and negative mood items used this same response format. However, if endorsing positive mood items is assigned a high score, then the scale is essentially a positive mood scale. Endorsing a positive mood item should result in a high value, and endorsing a negative mood item should yield a low value. This is what would be obtained if, for all negative mood items, responses of 7 were changed to 1, 6 to 2, and so forth. This type of transformation can be accomplished by creating a new score from the old score with the following formula: NEW = $(J + 1)$ − OLD, where NEW and OLD refer to the transformed and original scores, respectively, and J is the original number of response options. In the example presented, J would equal 7 and $(J + 1)$ would be 8. Subtracting a score of 7 from 8 would yield 1, subtracting 6 would yield 2, and so forth.

Some negative correlations among items may not be correctable by reverse scoring items. For example, reverse scoring a given item might eliminate some negative correlations but create others. This usually indicates that some of the items simply do not belong because they are not consistently related to other items. Any item that is positively correlated with some and negatively correlated with others in a homogeneous set should be eliminated if no pattern of reverse scoring items eliminates the negative correlations.

Item-Scale Correlations

If we want to arrive at a set of highly intercorrelated items, then each individual item should correlate substantially with the collection of remaining items. We can examine this property for each item by computing its *item-scale correlation.* There are two types of item-scale correlation. The corrected item-scale correlation correlates the item being evaluated with all the scale items, excluding itself, while the uncorrected item-scale correlation correlates the item in question with the entire set of candidate items, including itself. If there were 10 items being considered for a scale, the corrected item-scale correlation for any one of the 10 items would consist of its correlation with the other 9 items. The uncorrected correlation would consist of its correlation with all 10. In theory, the uncorrected value tells us how representative the item is of the whole scale. This is analogous, for example, to correlating one subset of an IQ test with the entire test to determine if the subscale is a suitable proxy. However, although an uncorrected item-total correlation makes good conceptual sense, the reality is that the item's inclusion in the "scale" can inflate the correlation coefficient. The fewer the number of items in the set, the bigger the difference that inclusion or exclusion of the item under scrutiny will make. In general, it is probably advisable to examine the corrected item-total correlation. An item with a high value for this correlation is more desirable than an item with a low value.

Item Variances

Another valuable attribute for a scale item is *relatively high variance.* To take an extreme case, if all individuals answer a given item identically, it will not discriminate at all among individuals with different levels of the construct being measured, and its variance will be 0. In contrast, if the development sample is diverse with respect to the attribute of interest, then the range of scores obtained for an item should be diverse as well. This implies a fairly high variance. Of course, increasing variance by adding to the error component is not desirable.

Item Means

A mean *close to the center of the range* of possible scores is also desirable. If, for example, the response options for each item ranged from 1, corresponding to "strongly disagree," to 7, for "strongly agree," an item mean near 4 would be ideal. If a mean were near one of the extremes of the range, then the item might fail to detect certain values of the construct. A piling up of scores at the value 7, for example, would suggest that the item was not worded strongly enough (i.e., that it was rare to find anyone who would disagree with it).

Generally, items with means too near to an extreme of the response range will have low variances, and those that vary over a narrow range will correlate poorly with other items. As stated previously, an item that does not vary cannot covary. Thus either a lopsided mean or a low variance for any reason will tend to reduce an item's correlation with other items. As a result, you can usually concentrate primarily on the pattern of correlations among items as a gauge of their potential value. Inspecting means and variances, however, is a useful double-check once a tentative selection of items has been made on the basis of the correlations.

Factor Analysis

A set of items is not necessarily a scale. Items may have no common underlying variable (as in an index or emergent variable) or may have several. Determining the nature of latent variables underlying an item set is critical. For example, an assumption underlying alpha is that the set of items is unidimensional. The best means of determining which groups of items, if any, constitute a unidimensional set is by factor analysis. This topic is sufficiently important to merit an entire chapter (see Chapter 6). Although factor analysis requires substantial sample sizes, so does scale development in general. If there are too few respondents for factor analysis, the entire scale development process may be compromised. Consequently, factor analysis of some sort should generally be a part of the scale development process at this stage.

Coefficient Alpha

One of the most important indicators of a scale's quality is the reliability coefficient, alpha. Virtually all the individual-item problems discussed thus far—a noncentral mean, poor variability, negative correlations among items, low item-scale correlations, and weak interitem correlations—will tend to reduce alpha. Therefore, after we have selected our items, weeding out the poor ones and retaining the good ones, alpha is one way of evaluating how

successful we have been. Alpha is an indication of the proportion of variance in the scale scores that is attributable to the true score. There are several options for computing alpha, differing in degree of automation. Some computer packages have item analysis programs that compute alpha. In SPSS, the Reliability procedure computes alpha for a full scale and for all $k - 1$ versions (i.e., every possible version with a single item removed). The program also provides corrected and uncorrected item-scale correlations. SAS includes alpha calculations as a feature of the correlation program, Proc Corr. By including the Alpha option – within Proc Corr, the variables listed in the accompanying Var (i.e., variable specification) statement will be treated as a scale, and alpha will be computed for the full set of items as well as all possible $k - 1$ item sets. Item-scale correlations are also provided.

Another option for computing alpha is to do so by hand. If variances for the individual items and for the scale as a whole are available, they can be plugged into the first formula for alpha discussed in Chapter 3. Or one can use the Spearman-Brown formula, which was also introduced in Chapter 3. This formula uses information available from a correlation matrix rather than variances as the basis for computing alpha. A shortcoming of this approach is that correlations are standardized covariances and standardizing the individual items might affect the value of alpha. If one adheres strictly to the model of parallel tests, then this is inconsequential because the correlations are assumed to be equal. However, they virtually never are exactly equal. The essentially tau-equivalent tests model does not require equal correlations among items, only equal covariances. Thus the proportion of each individual item's variance that is due to error is free to vary under that model. However, because the Spearman-Brown formula actually works with *average* interitem correlations, and one of the implications of the tau-equivalent model is that the average item-scale correlations are equal for each item, there is still no problem. Nonetheless, there can be small (but sometimes large) differences between the values of alpha obtained from covariance-based versus correlation-based computational methods. Because the covariance matrix uses the data in a purer form (without standardization), it is preferred and should generally be used.

Theoretically, alpha can take on values from 0.0 to 1.0, although it is unlikely that it will attain either of these extreme values. If alpha is negative, something is wrong. A likely problem is negative correlations (or covariances) among the items. If this occurs, try reverse scoring or deleting items as described earlier in this chapter. Nunnally (1978) suggests a value of .70 as a lower acceptable bound for alpha. It is not unusual to see published scales with lower alphas. Different methodologists and investigators begin to squirm at different levels of alpha. My personal comfort ranges for research scales are as follows: below .60, unacceptable; between .60 and .65, undesirable; between .65 and .70, minimally acceptable; between .70 and .80, respectable;

between .80 and .90, very good; much above .90, one should consider shortening the scale (see the following section on scale length). I should emphasize that these are *personal and subjective* groupings of alpha values. I cannot defend them on strictly rational grounds. However, they reflect my experience and seem to overlap substantially with other investigators' appraisals. The values I have suggested apply to *stable* alphas. During development, items are selected, either directly or indirectly, on the basis of their contribution to alpha. Some of the apparent covariation among items may be due to chance. Therefore, it is advisable during the development stage to strive for alphas that are a bit higher than you would like them to be. Then, if the alphas deteriorate somewhat when used in a new research context, they will still be acceptably high. As noted earlier, if the developmental sample is small, the investigator should be especially concerned that the initial alpha estimates obtained during scale development may not be stable. As we shall see, this is also the case when the number of items making up the scale is small.

A situation in which the suggested "comfort ranges" for alpha do not apply is when one is developing a scale that requires critical accuracy. Clinical situations are an example. The suggested guidelines are suitable for *research instruments* that will be used with *group data.* For example, a scale with an alpha of .85 is probably perfectly adequate for use in a study comparing groups with respect to the construct being measured. Individual assessment, especially when important decisions rest on that assessment, demands a much higher standard. Scales that are intended for individual diagnostic, employment, academic placement, or other important purposes should probably have considerably higher reliabilities, in the mid-.90s, for example.

In some situations, such as when a scale consists of a single item, it will be impossible to use alpha as the index of reliability. If possible, some reliability assessment should be made. Test-retest correlation may be the only option in the single-item instance. Although this index of reliability is imperfect, as discussed in Chapter 3, it is clearly better than no reliability assessment at all. A preferable alternative, if possible, would be to constitute the scale using more than a single item.

STEP 8: OPTIMIZE SCALE LENGTH

Effect of Scale Length on Reliability

At this stage of the scale development process, the investigator has a pool of items that demonstrates acceptable reliability. A scale's alpha is influenced

by two characteristics: the extent of covariation among the items and the number of items in the scale. For items that have item-scale correlations about equal to the *average* interitem correlation (i.e., items that are fairly typical), adding more will increase alpha and removing more will lower it. Generally, shorter scales are good because they place less of a burden on respondents. Longer scales, on the other hand, are good because they tend to be more reliable. Obviously, maximizing one of these assets reduces the other. Therefore, the scale developer should give some thought to the optimal trade-off between brevity and reliability.

If a scale's reliability is too low, then brevity is no virtue. Subjects may, indeed, be more willing to answer a 3-item than a 10-item scale. However, if the researcher cannot assign any meaning to the scores obtained from the shorter version, then nothing has been gained. Thus the issue of trading off reliability for brevity should be confined to situations when the researcher has "reliability to spare." When this is, in fact, the case, it may be appropriate to buy a shorter scale at the price of a bit less reliability.

Effects of Dropping "Bad" Items

Whether dropping "bad" items actually increases or slightly lowers alpha depends on just how poor the items are that will be dropped, and on the number of items in the scale. Consider the effect of more or fewer items that are equally "good" items—that is, that have comparable correlations with their counterparts: With fewer items, a greater change in alpha results from the addition or subtraction of each item. If the average interitem correlation among four items is .50, the alpha will equal .80. If there are only three items with an average interitem correlation of .50, alpha drops to .75. Five items with the same average correlation would have an alpha of .83. For 9-, 10-, and 11-item scales with average interitem correlations of .50, alphas would be .90, .91, and .92, respectively. In the latter instances, the alphas are not only higher, but they are much closer in value to one another.

If an item has a sufficiently lower-than-average correlation with the other items, dropping it will raise alpha. If its average correlation with the other items is only slightly below (or equal to, or above) the overall average, then retaining the item will increase alpha. I stated above that a 4-item scale would attain an alpha of .80, with an average interitem correlation of .50. How low would the average correlation of 1 item to the other 3 have to be for that item's elimination to help rather than hurt alpha? First, consider what the average interitem correlation would have to be for a 3-item scale to achieve an alpha of .80. It would need to be .57. So, after eliminating the worst of 4 items, the

remaining three would need an average interitem correlation of .57 for alpha to hold its value of .80. Three items whose average interitem correlation was lower than .57 would have a lower alpha than 4 items whose interitem correlations averaged .50. Assuming that the 3 best items of a 4-item scale had an average correlation among them of .57, the average correlation between the remaining (and thus worst) item and the other 3 would have to be lower than .43 for its elimination to actually increase alpha. (Having 3 items whose intercorrelations average .57 and 1 whose average correlation with the other 3 is .43 yields an overall average interitem correlation among the 4 of .50.) For any value larger than .43, having a fourth item does more good than lowering the average interitem correlation does harm. Thus the 1 "bad" item would have to be a fair bit worse than the other 3 (.57 − .43 = .14) to be worth eliminating.

Now, consider the situation when there is a 10-item scale with an alpha = .80. First of all, the average interitem correlation need only be about .29, illustrating the manner in which more items offset weaker correlations among them. For a 9-item scale to achieve the same alpha, the average interitem correlation would need to be about .31. A "bad" item would need to have an average interitem correlation with the remaining 9 items of about .20 or less in order for its inclusion as a tenth item to pull the overall average interitem correlation below .29. A failure to bring the average below this value would result in the item's inclusion benefiting alpha. The average interitem correlation difference between the 9 "good" items and the 1 "bad" item in this case is .31 − .20 = .11, a smaller difference than the one found in the 4-item example.

Tinkering With Scale Length

How does one go about "tinkering" with scale length in practice? Obviously, items that contribute least to the overall internal consistency should be the first to be considered for exclusion. These can be identified in a number of ways. The SPSS RELIABILITY procedure and the ALPHA option of PROC CORR in SAS show what the effect of omitting each item would be on the overall alpha. The items whose omission has the least negative or most positive effect on alpha is usually the best one to drop first. The item-scale correlations can also be used as a barometer of which items are expendable. Those with the lowest item-scale correlations should be eliminated first. SPSS also provides a squared multiple correlation for each item, obtained by regressing the item on all of the remaining items. This is an estimate of the item's *communality,* the extent to which the item shares variance with the other items. As with item-scale correlations, items with the lowest squared multiple correlations should be the prime candidates for exclusion. Generally,

these various indices of item quality converge. A poor item-scale correlation is typically accompanied by a low squared multiple correlation and a small loss, or even a gain, in alpha when the item is eliminated. Scale length affects the precision of alpha. In practice, a computed alpha is an estimate of reliability dependent on the appropriateness of the measurement assumptions to the actual data. It has already been noted that alpha increases when more items are included (unless they are relatively poor items). In addition, the *reliability of alpha as an estimate of reliability* increases with the number of items. This means that an alpha computed for a longer scale will have a narrower confidence interval around it than will an alpha computed for a shorter scale. Across administrations, a longer scale will yield more similar values for alpha than will a shorter one. This fact should be considered in deciding how long or short to make a scale during development.

Finally, it is important to remember that a margin of safety should be built into alpha when trying to optimize scale length. Alpha may decrease somewhat when the scale is administered to a sample other than the one used for its development.

Split Samples

If the development sample is sufficiently large, it may be possible to split it into two subsamples. One can serve as the primary development sample, and the other can be used to cross-check the findings. So, for example, data from the first subsample can be used to compute alphas, evaluate items, tinker with scale length, and arrive at a final version of the scale that seems optimal. The second subsample can then be used to replicate these findings. The choice of items to retain will not have been based at all on the second subsample. Thus alphas and other statistics computed for this group would not manifest the chance effects, such as alpha inflation, that were discussed earlier. If the alphas remain fairly constant across the two subsamples, you can be more comfortable assuming that these values are not distorted by chance. Of course, the two subsamples are likely to be much more similar than two totally different samples. The subsamples, divided randomly from the entire development sample, are likely to represent the same population; in contrast, an entirely new sample might represent a slightly different population. Also, data collection periods for the two subsamples are not separated by time, whereas a development sample and a totally separate sample almost always are. Furthermore, any special conditions that may have applied to data collection for one subsample would apply equally to the other. Examples of special conditions include exposure to specific research personnel, physical settings, and clarity of questionnaire printing. Also, the two subsamples may be the only

two groups to complete the scale items together with all of the items from the original pool that were eventually rejected. If rejected items exercised any effects on the responses to the scale items, these would be comparable for both subsamples.

Despite the unique similarity of the resultant subsamples, replicating findings by splitting the developmental sample provides valuable information about scale stability. The two subsamples differ in one key aspect: In the case of the first subsample on whose data item selection was based, the opportunity existed for unstable, chance factors to be confused with reliable covariation among items. No such opportunity for systematically attributing chance results to reliability exist for the second group because its data did not influence item selection. This crucial difference is sufficient reason to value the information that sample splitting at this stage of scale development can offer.

The most obvious way to split a sufficiently large sample is to halve it. However, if the sample is too small to yield adequately large halves, you can split it unevenly. The larger subsample can be used for the more crucial process of item evaluation and scale construction and the smaller for cross-validation.

EXERCISES

Assume that you are developing a fear-of-snakes measure, with a 6-choice Likert response format, using 300 subjects. Although more items would be desirable for actual scale development, for these exercises:

1. Generate a pool of 10 Likert-format items.
2. Estimate, for each item you have written, what Likert scale values would be endorsed by the "average person" (i.e., neither a snake phobic nor a snake charmer).
3. Pick an item from the pool that you suspect might elicit an extreme response from an average person and rewrite it to elicit a more moderate response.
4. Generate another 10 Likert items to tap a construct *other than* fear of snakes. Randomly mix these items with the original 10 and ask a few of your friends to indicate what they think each of the items is intended to measure.
5. Using either fear of snakes or the construct underlying your second pool of 10 items, list directly observable behaviors that could be used to validate a scale measuring that construct and explain how you could use behavioral data for validation.

6. What would the alpha for the scale be if your 10 fear-of-snakes items had an average interitem correlation of .30?[1]
7. How could you use split samples to estimate and cross-validate the scale's coefficient alpha?

NOTE

1. For Exercise 6, alpha = $[10 \times .30]/[1 + (9 \times .30)] = .81$.

6

Factor Analysis

In Chapter 2, when discussing different theoretical models that could describe the relationship of a scale's items to the latent variable, I mentioned the general factor model. That model does not assume that only one latent variable is the source of all covariation among the items. Instead, the model allows multiple latent variables to serve as causes of variation in a set of items.

To illustrate how more than one latent variable might underlie a set of items, I will describe a specific, albeit hypothetical, situation. Many constructs of interest to social and behavioral scientists can be operationalized at multiple levels of specificity. The terms *psychological adjustment, affect, negative affect, anxiety,* and *test anxiety* are examples of hierarchical phenomena. Each term could subsume those that follow it in the list, and it might be possible to develop measures at each level of specificity. Presumably, differently worded items with different time frames and response options could tap either a specific, middling, or general level of this continuum. Hopefully, a scale developer would select item wordings that corresponded to the intended level of variable specificity. Factor analysis then could be used to assess whether that selection process succeeded.

To make this example more specific, consider a set of 25 items that all pertain to affect. Our concern is whether these items should make up one general scale or many more specific scales. Do all 25 items belong together? Or, is it more appropriate to have separate scales for different affective states, such as depression, euphoria, hostility, anxiety, and so on? Maybe it would be even better to split the positive and negative affect items (e.g., "happy" versus "sad" for depression or "tense" versus "calm" for anxiety) into separate scales. How do we know what is most appropriate for the items at hand? Essentially, the question is, does a set of items asking about several affective states have one or many latent variables underlying it?

Attempting to answer these questions using only the methods other than factor analysis discussed in the preceding chapters would be daunting. We could compute alpha on the entire set of mood items. Alpha would tell us something about how much variance a group of items had in common. If alpha were low, we might search for subsets of items that correlate strongly with each other. For example, we might suspect that positive and negative affect items do not correlate with one another and that combining them was

lowering alpha. The alphas for these more homogeneous (all positive or all negative affect) subsets of items should be higher. We might then speculate that even more homogeneous subsets (e.g., separating anxiety from depression in addition to positive from negative) should have still higher alphas. However, at some point, we might also worry that these more specific and homogeneous scales would correlate strongly with each other because they were merely tapping different aspects of the same affective state. This would suggest that their items belonged in the same rather than in separate scales.

It should be emphasized that a relatively high alpha is no guarantee that all the items reflect the influence of a single latent variable. If a scale consisted of 25 items, 12 reflecting primarily one latent variable and the remaining 13 primarily another, the correlation matrix for all the items should have some high and some low values. Correlations based on two items representing primarily the same latent variable should be high, and those based on items primarily influenced by different latent variables should be relatively low. However, the *average* interitem correlation might be high enough to yield a respectable alpha for a 25-item scale. For example, to yield an alpha of .80, the average interitem correlation needs to be only .14.

Factor analysis, the topic of this chapter, is a useful analytic tool that can tell us, in a way that reliability coefficients cannot, about important properties of a scale. It can help us determine *empirically* how many constructs, or latent variables, or factors underlie a set of items.

AN OVERVIEW OF FACTOR ANALYSIS

Factor analysis serves several related purposes. One of its primary functions, as just noted, is to help an investigator determine *how many latent variables* underlie a set of items. Thus, in the case of the 25 affect items, factor analysis could help the investigator determine whether one broad or several more specific constructs were needed to characterize the item set. Factor analysis also can provide a means of explaining variation among relatively many original variables (e.g., 25 items) using relatively few newly created variables (i.e., the factors). This amounts to *condensing* information so that variation can be accounted for by using a smaller number of variables. For example, instead of needing 25 scores to describe how respondents answered the items, it might be possible to compute fewer scores (perhaps even one score) based on combining items. A third purpose of factor analysis is to *define the substantive content or meaning of the factors* (i.e., latent variables) that account for the variation among a larger set of items. This is accomplished by identifying

groups of items that covary with one another and appear to define meaningful underlying latent variables. If, say, two factors emerged from an analysis of the 25 affect items, the individual items making up those factor groupings could provide a clue about the underlying latent variables represented by the factors.

The following sections present a conceptual summary of factor analysis. Readers who want a more thorough treatment of factor analysis should consult a text devoted to the topic, such as Cureton (1983), Gorsuch (1983), Harman (1976), or McDonald (1984).

Examples of Methods Analogous to Factor Analytic Concepts

To get an intuitive sense of what factor analysis does, we can consider two less formal but roughly analogous processes with which we may be more familiar. The first of these processes is sometimes used in human resources management to identify common themes among seemingly diverse specific issues that may concern team members or co-workers.

Example 1

Assume that a small, new company wants to identify what characteristics its employees believe are important for their co-workers to have. They believe that identifying and rewarding widely valued characteristics will play an important part in cultivating a harmonious and cooperative work environment. The company hires a human resources specialist to assist them. This person, whom we will call Jim, gathers the company's 10 employees together and explains that he would like them to consider what characteristics of their fellow employees they regard as important across the range of interactions they might have on the job, from developing proposals and reports together, to interacting with potential clients together, to sharing a table in the cafeteria—the full range of interactions employees might have. Jim suggests that, to begin the process, the employees each write on separate pieces of paper as many important characteristics as they can identify.

After several minutes during which employees write down their ideas, Jim asks for a volunteer to read one of his or her ideas to the group. Alice says that one characteristic she wrote down was "willing to share ideas." Jim thanks her and asks that she tape the paper with that idea on it to the wall. Another employee, Bill, reads one of his characteristics: "a sense of humor." This too is taped to the wall. The process continues with each individual employee stating each of the characteristics he or she had written down. In this way, people individually name a variety of characteristics that they personally consider

important in co-workers. After doing so, they tape the sheet of paper naming each characteristic to the wall. Among the characteristics listed are the following:

willing to share ideas
sense of humor
always has the right tools for the job
smart
isn't sloppy
hard worker
job at hand
thinks logically
comes through in a pinch
prepares for tasks
makes good impression with clients
doesn't try to get all the credit
fun
has a nice car
has a lot of experience in this type of work
is friendly
can be counted on
pays attention to details
has a mind like a steel trap
is outgoing
knows a lot of potential clients
is reliable
has character
is well educated
is trustworthy
knows how to dress
is a good storyteller
is intelligent
is a person of faith
is willing to work long hours if that's what it takes to get the job done

This process continues for some time, and soon the wall is covered with more than 30 slips of paper, each naming a characteristic that an employee thought was important. Next, Jim asks if people see any characteristics that they think go together. Katherine points out that "smart" and "is intelligent" are the same. Jim takes the sheet of paper on which "is intelligent" is written and moves it next to "smart." Frank suggests that "is well educated" should also be part of that group. Several other characteristics are added to the same group of statements. Then Carla observes that "is friendly" and "makes a good impression with clients" are similar to each other and different from the group of statements already formed. She suggests that those two characteristics should be put together in a new group. Then, "fun" is also added to this second group. "Isn't sloppy" and "knows how to dress" form the kernel of a third group until one employee says she thinks "isn't sloppy" would go better with "prepares for tasks" than with "knows how to dress." This process continues until Jim and the employees have formed several clusters of statements. Virtually every characteristic described gets placed into some group.

Jim then asks people to name—with a word or short descriptive phrase—each group of statements. The various groups of items are labeled "Intelligence,"

"Appearance," "Conscientiousness," "Personality," "Dependability," and so on. Presumably, each group of statements represents a key concept related to employees' perceptions of one another's characteristics.

Example 2

Several years later, the company decides to repeat the exercise. The managers suspect that things have changed enough that not all of the categories originally identified may still be relevant. Jim, the human resources facilitator, is unavailable. Carol, one of the company's executives, decides that an easier way of getting at similar information might be to develop a questionnaire that has statements like the ones people came up with in the earlier exercise. Employees would be asked to indicate how important they felt each characteristic was, using "not at all," "somewhat," and "very" as their response options. These questionnaires were administered to the employees, who now numbered nearly 150. When Carol got the questionnaires back, she looked them over to see which things were most important. One thing she noticed was that different things were important to different people, but certain characteristics tended to be rated similarly to one another. For example, people who thought "pays attention to details" was important were likely to consider "prepares for tasks" important as well. People who did not consider one of these items important typically did not consider the other important either. Carol mulled over the pile of questionnaires and thought about how to make sense of them. She remembered how, during the original exercise conducted several years earlier with the slips of paper on the wall, there seemed to be more groups than were really needed. Some of the individual statements were fairly worthless, she thought, and sometimes if the same person had more than one of these worthless statements, a whole worthless category would result. She wondered if there was a way of determining how many categories it would take to distill most of what the employees thought about their co-workers. As an exercise, she tried looking for other sets of items, like the two she had noticed already, that tended to be endorsed similarly across employees. In essence, she looked for groupings of similar items based not on just their content but also on how similarly they were evaluated by employees. This took a great deal of time and Carol was not really sure if she was picking up on all the important clusters of statements, but she felt able to garner some interesting ideas in this way from the questionnaires.

Shortcomings of These Methods

Both of these examples are conceptually analogous to factor analysis but with certain important differences. In both cases, the result is a reorganization

of a substantial amount of specific information into a more manageable set of more general but meaningful categories. Presumably, each of these reclassifications resulted in a few ideas that captured much of what the many individual statements covered. Both approaches had fairly obvious shortcomings, however.

In the first example, there was little control over the quality of the statements generated. Some people are more extroverted than others and may end up generating more statements. It is not always the case, however, that the most extroverted group members will be the most insightful. For this and other reasons, this process often leads to statements that are ambiguous, irrelevant, or downright silly. Depending on the dynamics of the group involved in the exercise, it may be difficult to dismiss such statements without offending their authors. Consequently, such statements end up being given as much credibility as better statements. Even if the statements are generally relevant and avoid silliness, some will be more germane to the issues at hand than are others. But, again, all may tend to be treated more or less equally. If several silly but similar items are offered, they are likely to constitute a category based merely on their similarity. Categories may be prioritized, but this is usually done by consensus and, depending on who generated the statements, there may be reluctance to brand certain categories as trivial. My experience with exercises of this sort suggests, furthermore, that there is a strong tendency to place every statement in *some* category. Having several neat categories and then one or two orphaned statements seems to leave people with a lack of closure, so they give orphaned statements a home, even if the fit is not immediately evident. A final limitation is that, although one can identify which specific statements exemplify a category, it is not necessarily obvious which are better or worse exemplars.

The second example avoided some of these shortcomings. Carol could weed out items that struck her as irrelevant, although this places quite a burden on her judgment. At least the process of endorsing items was somewhat more democratic. Every person got to evaluate every item without risk of alienating a co-worker. The groupings were determined not by a mere sense of apparent statement similarity but by evidence that people reacted to similarly grouped items in a common way. That is, the similarity was a characteristic of the items (certain sets of which seemed to elicit similar perceptions), not the respondents (who varied in their responses to any specific item). Seeing one item in a group as unimportant implies a substantial likelihood of the same individual seeing the other items in the same group as unimportant. A different employee might see those same items as consistently important. The critical issue is that, whatever an individual's assessment of items' importance, it tended to be consistent across statements within a group. In fact, that was the basis on

which Carol constituted the groups. Doing this by visual inspection for 50 questionnaires would be fairly daunting, and thus it is likely that Carol's categorization system was not the most efficient possible method. How much consistency was required for items to be considered a group? How many instances of a single employee giving divergent assessments (i.e., an agreement and a disagreement of importance) for two items in the same potential cluster would Carol tolerate?

A CONCEPTUAL DESCRIPTION OF FACTOR ANALYSIS

Factor analysis is a category of procedures that accomplishes the same type of classification as the methods described above, but it does so in accordance with a more structured set of operations and provides more explicit information that the data analyst can use to make judgments. Like the methods just described, factor analysis identifies categories of similar statements. The factor analyst's first task is to determine how many categories are sufficient to capture the bulk of the information contained in the original set of statements.

Extracting Factors

In essence, factor analysis begins with the premise that one big category containing all of the items is all that is needed (i.e., that one concept or category is sufficient to account for the pattern of responses). It then assesses how much of the association among individual items that single concept can explain. The analysis then performs a check to see how well the single-concept premise has fared. If it appears that one concept or category has not done an adequate job of accounting for covariation among the items, the factor analysis rejects the initial premise. It then identifies a second concept (i.e., latent variable or factor) that explains some of the remaining covariation among items. This continues until the amount of covariation that the set of factors has not accounted for is acceptably small.

The First Factor

How is this accomplished? The process begins with a correlation matrix for all of the individual items. Using this matrix as a starting point, factor analysis examines the patterns of covariation represented by the correlations among items. What follows is a conceptual description. Certain mathematical details are omitted in the interest of clarity, so this should not be taken literally as the set of operations underlying computer-generated factor analyses.

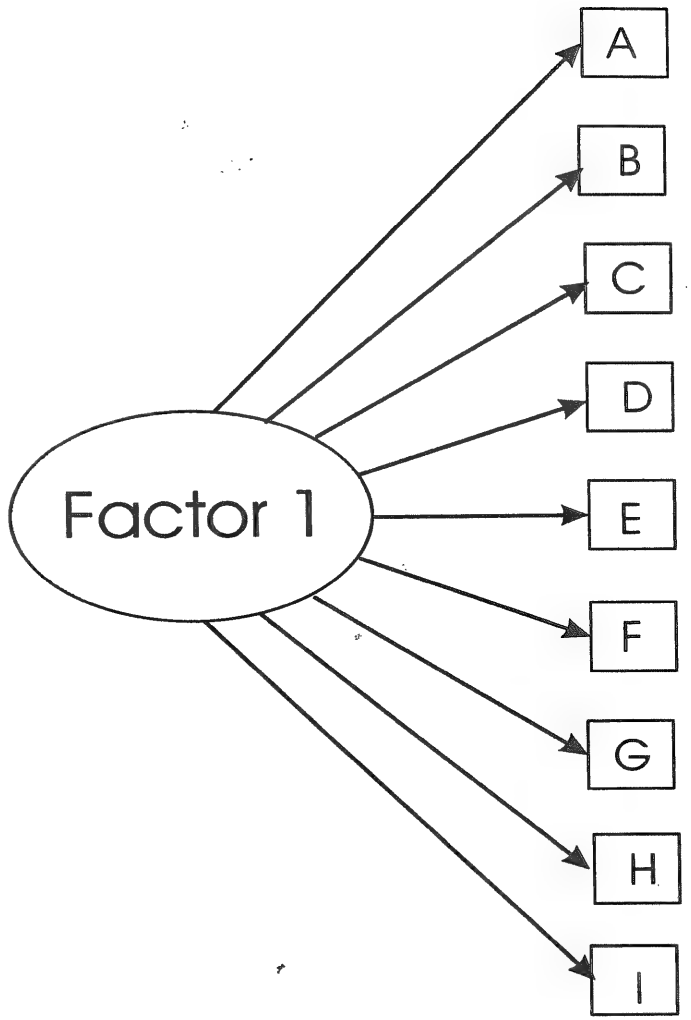

Figure 6.1 A single factor model

As stated earlier, the process involves the initial premise of a single concept that can adequately account for the pattern of correlations among the items. This amounts to a provisional assertion that a model that has a single latent variable (i.e., a single factor), with a separate path emanating from it to each

of the items, is an accurate representation of causal relationships. This further implies that such a model can account for the correlations among the items. To test this assumption conceptually, the factor analysis program must determine the correlation of each item with the factor representing the single latent variable, and then see if the observed correlations between items can be recreated by appropriately multiplying the paths linking each pair of variables via the factor. But how can the program compute correlations between observed item responses and a factor representing a latent variable that has not been directly observed or measured?

One approach is to posit that the sum of all the item responses is a reasonable numerical estimate of the one, all-encompassing latent variable that is assumed to account for interitem correlations. In essence, this overall sum is an estimate of the latent variable's "score." Because the actual scores for all items are presumed to be determined by one latent variable, a quantity combining information from all items (i.e., an overall sum) is a reasonable estimate of that latent variable's numerical value. It is fairly simple to add the individual item scores together into a total score and to compute item-total correlations for each separate item with the total of all items. These item-total correlations serve as proxies for the correlations between the observed items and the unobserved latent variable (i.e., the causal pathways from the latent variable to the individual items). With values thus assigned to those causal pathways, one then can compute projected interitem correlations based on this one-factor model. These model-derived correlations are projections of what the actual interitem correlations should be if the premise of only one underlying variable is correct. The legitimacy of the premise can be assessed by comparing the projected correlations to the actual correlations. This amounts to subtracting each projected correlation from the corresponding actual correlation based on the original data. A substantial discrepancy between actual and projected correlations indicates that the single-factor model is not adequate, that there is still some unaccounted-for covariation among the items.

Consider this sequence for a single pair of items, A and B, that are part of a larger set. First, the whole set of items, including A and B, would be added together to get a summary score. Then, correlations of A with that total score and B with that total score would be computed. These two item-total correlations are assumed to represent the correlations of A and of B with the factor, which corresponds to the underlying latent variable. If the premise of a single underlying latent variable is correct, then a path diagram involving A, B, and the factor would have paths (a and b in Figure 6.2) from the latter to each of the former. The value of these paths would be the item-total correlations just described. Based on this path diagram, the correlation between A and B should be the product of those two paths. Computing this proposed correlation

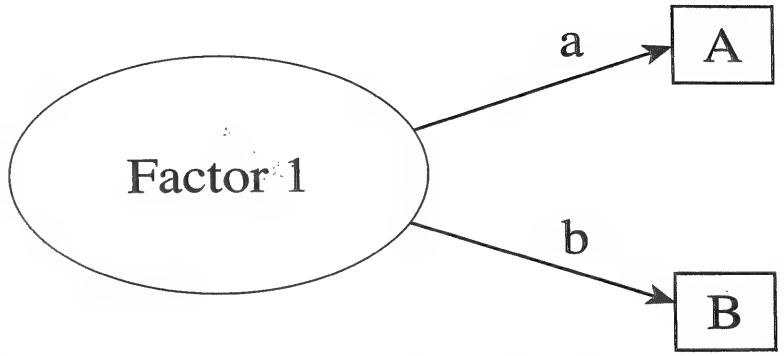

Figure 6.2 A simplified single-factor model involving only two items

between A and B entails simple multiplication. Once computed, the proposed correlation can be compared to the actual correlation between A and B. The proposed correlation can be subtracted from the actual correlation to yield a residual correlation. A substantial residual correlation would indicate that invoking a single underlying latent variable as the sole cause of covariation between A and B is not adequate.

Operations performed on the whole correlation matrix simultaneously do this for each possible pairing of items. Rather than ending with a single residual correlation, one computes an entire matrix of residual correlations (called, appropriately, a *residual matrix*), each representing the amount of covariation between a particular pair of items that exists above and beyond the covariation that a single latent variable could explain.

Subsequent Factors

It is now possible to treat this residual matrix in the same way the original correlation matrix was treated, extracting a second factor corresponding to a new latent variable. Once again, correlations between the items and that second latent variable (i.e., Factor 2) can be computed and, based on those correlations, a matrix of proposed correlations can be generated. Those proposed correlations represent the extent of correlation that should remain among items after the second factor has been taken into consideration. If the second factor captured all of the covariation left over after extracting the first factor, then these projected values should be comparable to the values that were in the residual matrix mentioned above. If not, further factors may be needed to account for the remaining covariation not yet ascribed to a factor.

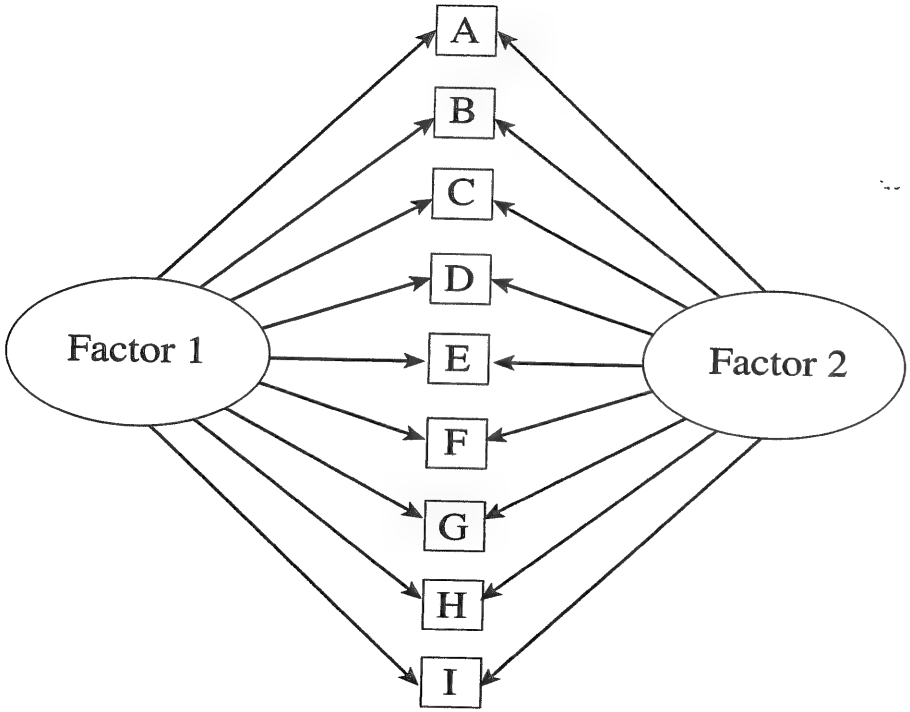

Figure 6.3 A two-factor model

This process can proceed, with each successive factor being extracted from the residual matrix that resulted from the preceding iteration, until a matrix is achieved that contains only acceptably small residual correlations. At this point, one can determine that essentially all of the important covariation has been accounted for and that no further factors are needed. It is possible to continue the process until a residual matrix consisting entirely of zeros is obtained. This will occur when the number of factors extracted equals the number of items in the factor analysis. Stated differently, a set of k factors will always be able to explain all of the covariation among a set of k items.

Deciding How Many Factors to Extract

Determining how many factors to extract can be a knotty issue (e.g., Zwick & Velicer, 1986). Of course, a major motivation for conducting factor analysis is to move from a large set of variables (the items) to a smaller set (the

factors) that does a reasonable job of capturing the original information, that is, to condense information. Determining what is "a reasonable job" can be approached in several ways.

Some factor analytic methods, such as those based on maximum likelihood estimation and confirmatory factor analytic procedures (which we will examine subsequently) based on structural equation modeling approaches, use a statistical criterion. In this context, the term *statistical criterion* refers to the fact that inferential methods are used to determine whether the likelihood of a particular result is sufficiently small to rule out its chance occurrence. This amounts to performing a test to see if, after extracting each successive factor, the remaining residuals contain an amount of covariation statistically greater than zero. If they do, the process continues until that is no longer the case. The reliance on a statistical criterion rather than a subjective judgment is an appealing feature of these approaches. However, in scale development, it may not correspond to the goal at hand, which is to identify a small set of factors that can account for the important covariation among the items. Statistically based methods seek an *exhaustive* account of the factors underlying a set of items. If some source of covariation exists that has not been accounted for by any of the factors yet extracted, such programs push further. What the scale developer often is after is a *parsimonious* account of the factors. That is, in the course of scale development, we often want to know about the few most influential sources of variation underlying a set of items, not every possible source we can ferret out. When developing a scale, typically one generates a longer list of items than is expected to find its way into the final instrument. Items that do not contribute to the major identifiable factors may end up being trimmed. Our goal is to identify relatively few items that are strongly related to a small number of latent variables. Although a skilled data analyst can achieve this goal by means of factor analytic methods using a statistical criterion, a less experienced investigator, paradoxically, might do better using other more subjective but potentially less cryptic guidelines.

These relatively subjective guidelines are often based on the proportion of total variance among the original items that a group of factors can explain. This is essentially the same basis used by the statistically based methods. In the case of nonstatistical (i.e., not based on likelihood) criteria, however, the *data analyst* assesses the amount of information each successive factor contains and judges when a point of diminishing returns has been reached. This is roughly analogous to interpreting correlations (such as reliability coefficients) on the basis of their magnitude, a subjective criterion, rather than their p value, a statistical criterion. Two widely used nonstatistical guidelines for judging when enough factors have been extracted are the eigenvalue rule (Kaiser, 1960) and the scree test (Cattell, 1966).

An *eigenvalue* represents the amount of information captured by a factor. For certain types of factor analytic methods (namely, principal components analysis, discussed later in this chapter), the total amount of information in a set of items is equal to the number of items. Thus, in an analysis of 25 items, there would be 25 units of information. Each factor's eigenvalue corresponds to some portion of those units. For example, in the case of a 25-item analysis, a factor with an eigenvalue of 5.0 would account for 20% (5/25) of the total information; one with an eigenvalue of 2.5 would account for 10%, and so on. A consequence of this relationship between how information is quantified and the number of items in the analysis is that an eigenvalue of 1.0 corresponds to $1/k$ of the total variance among a set of items. Stated differently, a factor (assuming principal components analysis) that achieves an eigenvalue of 1.0 contains the same proportion of total information as does the typical single item. Consequently, if a goal of factor analysis is to arrive at a smaller number of variables that substantially capture the information contained in the original set of variables, the factors should be more information-laden than the original items. Accordingly, the *eigenvalue rule* (Kaiser, 1960) asserts that factors with eigenvalues less than 1.0 (and thus containing *less* information than the average item) should *not* be retained. Although the rationale for excluding such factors makes sense, what about factors that are only slightly above 1.0? Does a factor that explains 1% more information than the typical item really offer the sort of condensation of information we are after? Oftentimes, the answer is no, suggesting that the eigenvalue rule may be too generous a basis for retaining factors. I believe this is generally the case in scale development based on classical methods.

The *scree test* (Cattell, 1966) is also based on eigenvalues but uses their relative rather than absolute values as a criterion. It is based on a plot of the eigenvalues associated with successive factors. Because each factor after the first is extracted from a matrix that is a residual of the previous factor's extraction (as described earlier), the amount of information in each successive factor is less than in its predecessors. Cattell suggested that the "right" number of factors can be determined by looking at the drop in amount of information (and thus, eigenvalue magnitude) across successive factors. When plotted, this information will have a shape characterized by a predominantly vertical portion on the left (representing large eigenvalues) transitioning to a relatively horizontal portion on the right (corresponding to small eigenvalues). He regarded the factors corresponding to the right-side, horizontal portion of the plot as expendable. In lay terms, *scree* describes the rubble that collects on the ground following a landslide. This term, then, implies that the vertical portion of the plot is where the substantial factors are located while the horizontal portion is the scree, or rubble, that should be discarded. Ideally, the progression of

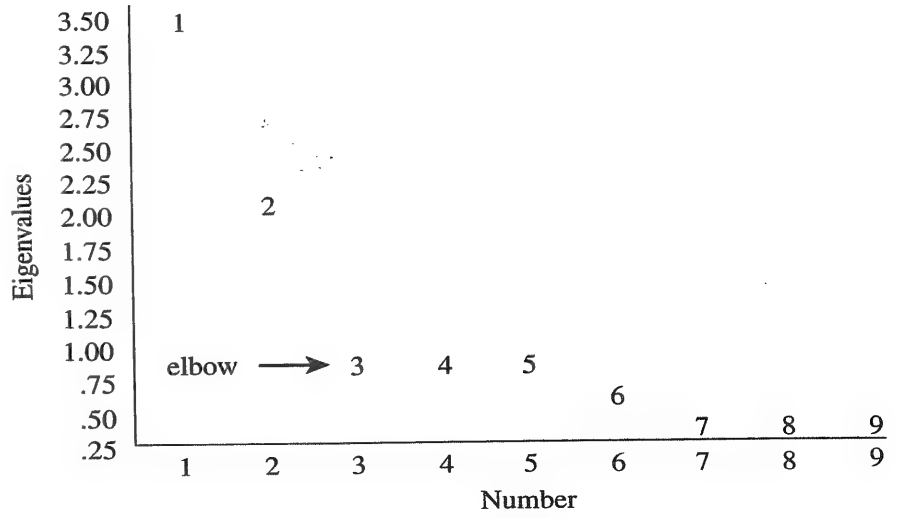

Figure 6.4 A scree plot with a distinct elbow

factors will have a point at which the information drops off suddenly, with an abrupt transition from vertical to horizontal and a clear "elbow."

Cattell's criterion calls for retaining those factors that lie above the elbow of the plot. Sometimes, the transition is not abrupt but gradual, with a gentle curve made up of several factors lying between the vertical and horizontal regions of the plot. In such cases, applying Cattell's scree test can be tricky and involves even greater reliance on subjective criteria, such as factor interpretability. A factor is considered interpretable to the extent that the items associated with it appear similar to one another and make theoretical and logical sense as indicators of a coherent construct.

Rotating Factors

The purpose of factor extraction is merely to determine the appropriate number of factors to examine. Putting information into the most understandable form is not its intent. The raw, unrotated factors are rather meaningless mathematical abstractions. As a rough analogy, imagine that I have been asked to describe the height of all the people in a room. I decide to do this by arbitrarily selecting a person, Joe, at random, measuring Joe's height, and describing everyone else as so many inches taller or shorter than the reference

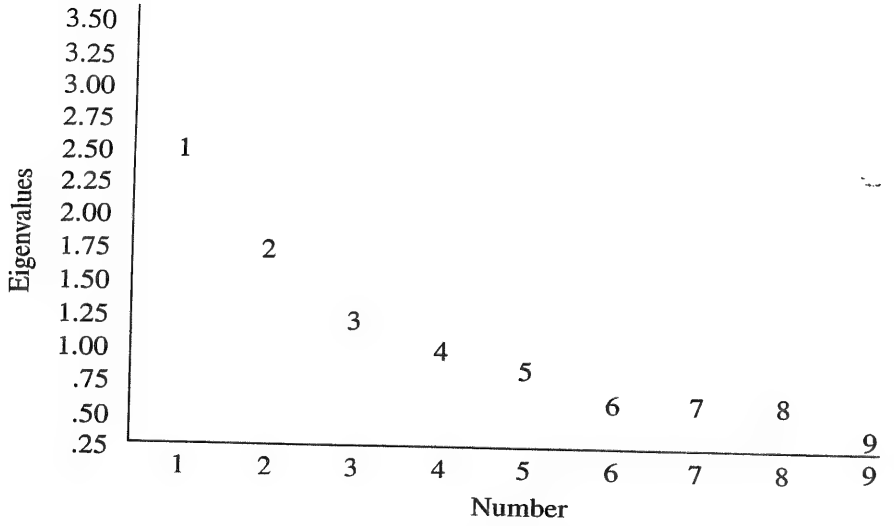

Figure 6.5 A scree plot without a distinct elbow

individual. So, one person might be "Joe plus 3 inches" and another, "Joe minus 2 inches." In such an instance, all the information about height is available in my presentation of the data, but it has not been organized in the most informative way. It would be easier for people to interpret my data if I transformed them into a more readily understandable form, such as the height of each individual in the room, expressed in feet and inches. Factor rotation is analogous to this transformation in that it presents data already available in a way that is easier to understand.

Before trying to interpret factors—to ascertain what the constructs or latent variables corresponding to factors are, based on the items identified with each factor—it is usually necessary to perform a factor rotation. Factor rotation increases interpretability by identifying clusters of variables that can be characterized predominantly in terms of a single latent variable, that is, items that are similar in that they all have a strong association with (and thus are largely determined by) only one and the same factor. Rotation and the greater interpretability that results from it are accomplished not by changing the items or the relationships among them but by selecting vantage points from which to describe them.

The patterns of intercorrelations among a set of items are analogous to physical locations in space. The more strongly two items are correlated, the

closer two markers representing those items could be placed to each other. If we did this for many items, the physical locations of their markers would take on a pattern representing the patterns of correlations among the variables. (This is easiest to visualize if we limit ourselves to two dimensions.) Imagining physical objects whose locations are determined by underlying rules is thus another way of thinking about items whose associations are determined by underlying causal variables.

Rotation Analogy 1

How does rotation allow us to see a pattern among variables that was always there but was not apparent? As an analogy, consider a very well-organized graveyard, such as Arlington National Cemetery, where uniform markers are neatly arranged. When viewed from a distance, it is possible to stand in certain locations such that the arrangement of graves in orderly rows and columns is completely obscured. If the line of sight does not follow along any of the natural axes of linear arrangements, the grave markers appear to be placed randomly. Changing one's vantage point, however, can reveal the underlying order. Stepping a few feet right or left can result in the line of sight falling along an alignment of markers and revealing their orderliness. Each marker, it is now apparent, shares a row (and a column) with other markers. All the markers in a given row thus have something in common—they possess a shared attribute (membership in the same row) that had not been evident from the earlier vantage point. Factor rotation is analogous in that it attempts to provide a "vantage point" from which the data's organizational structure—the ways in which items share certain characteristics—becomes apparent.

It is worth noting that, with the right number of perpendicular reference lines, one can locate objects no matter how those reference lines are oriented. A two-dimensional example, such as Arlington Cemetery (ignoring hills and valleys for the moment) illustrates this. I could draw a straight line through the cemetery at any orientation and then position a second line perpendicular to the first. With those two lines, I could specify the location of any grave marker: I could say, "Walk along Axis A for 50 yards, turn exactly 90° right (thus facing in a direction of travel parallel to Axis B), and proceed an additional 10 yards." This would place you at a specific location. I could get you to the same location using appropriately modified instructions based on any two perpendicular lines drawn through the cemetery. So, the orientation of the lines is arbitrary with respect to their adequacy of describing a specific location. Any set of two perpendicular lines has the same informational utility in locating a specific spot as any other set. Of course, this presumes having the

correct number of lines. For this example, I conceptualized the cemetery as essentially a two-dimensional space. Accordingly, two lines are sufficient and necessary to identify all possible locations in the cemetery. If I had just one orientation line, only by chance could a direction based on a position on that single line get you to the intended location. Factor rotation is a means of orienting the correct number (determined during the factor extraction process) of "lines" in a way that is most telling.

The operational definition of "most telling" concerns identifying inherent similarities in the items (analogous to the grave markers sharing a common row) and orienting the reference lines so that they coincide with a progression along that dimension of similarity (analogous to having an orientation line track straight along a row of markers, from the first in a row to the last in the same row). When this is achieved, progression along the dimension of similarity (e.g., along the rows of markers) can be adequately described by specifying locations along only one line instead of two. Although the cemetery has two important dimensions, this orientation allows us to describe a salient characteristic of the markers with only one variable—position along a row.

Analogies are often imperfect. In this case, it may seem that little has been gained by being able to characterize position along a row without regard for position along a column. The critical point is that orienting the reference lines appropriately *makes it possible* to summarize a feature of the graves by using a single value. The following analogy, while also imperfect, may make the utility of conveniently summarizing a single dimension somewhat clearer.

Rotation Analogy 2

Some cemeteries (to continue with what I hope isn't too morbid an example) are arranged chronologically from earlier to later burials. Some older European cemeteries are arranged with respect to status, with the most prominent and pious of the deceased being placed closest to the wall of an adjacent cathedral, for example. It is possible to imagine a cemetery arranged on the basis of both these criteria, with a date of burial varying from earlier to later along an orientation running parallel to the cathedral wall and prestige varying from more to less prestigious along an orientation running perpendicular to it. As before, any location in the cemetery can be specified by reference to any two perpendicular lines. But, note that I can summarize a lot of information about one common attribute of the grave plots using just one line (progressing from earlier to later burials) or the other line (progressing from less to more prestigious grave occupants) if I arrange them so that one is parallel and the other perpendicular to the cathedral. One dimension of similarity among the graves—prestige—falls along the dimension specified by one of

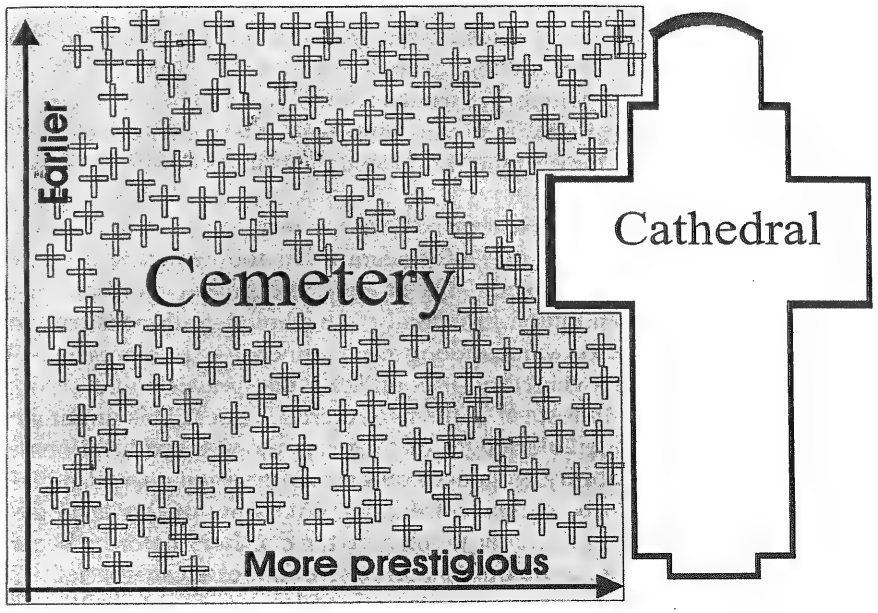

Figure 6.6 An imaginary cemetery with graves arranged according to the date of burial and prestige of the deceased

the reference lines. The other dimension of similarity—how long ago the grave was occupied—falls along the other (see Figure 6.6).

If a group of tourists wanted to know where the tombs of the prominent citizens the city has produced over the years are likely to be located, I could direct them to follow a line headed directly toward the side of the cathedral and tell them that as they walk along that line, the prominence of the interred will progressively increase. Where along the cathedral wall the line terminated would be irrelevant as far as prestige was concerned. I could similarly instruct tourists looking for the oldest graves to follow a line parallel to the cathedral's longer axis. Walking along such a line, they would encounter successively older graves. How close or far from the cathedral wall this line was would be irrelevant with respect to antiquity of the tombs.

These are useful locations for my two reference lines because they allow me to summarize information about one or the other of two variables determining the organizational structure of the cemetery with reference to only a single

value (i.e., location along one or the other axis). Rotating both axes, say, 43.5° clockwise, would result in a far less useful set of reference lines. Someone exploring the cemetery by following these displaced axes might not as easily recognize the scheme on which the whole collection of tombs is organized. The axes could be used to specify any location with as much accuracy as the earlier pair of lines, but they would not provide any insight into the way the locations were organized.

This cemetery analogy differs from the factor analysis of actual items in an important way, in that all possible locations in the cemetery as I have described it thus far are occupied. That is, there are graves representing all degrees of prestige and antiquity. Essentially, the whole two-dimensional grid is filled. The full range of values along each dimension is represented by graves. Also, in an actual cemetery, it is unlikely that graves could be organized in perfect order with respect to the two reference lines. This might work initially and for a considerable length of time. As more grave locations are claimed, however, the ability to adhere strictly to the oldness and prestige placement criteria would be compromised, a shortcoming of this analogy as presented thus far. It might be better to imagine the cemetery before it became crowded, when there were more options for where to locate each burial.

A less densely populated cemetery also strengthens our analogy to factor rotation in scale development because in the latter case, we would not typically have items representing all possible degrees of the two dimensions in question. When we write items, we make a conscious attempt not to have them tap more than one variable of interest. If an item appears to tap more than one underlying variable, we discard it. Consequently, in the two-variable case, we would try to have strong, unambiguous items for each of the variables we hoped to measure. This would be analogous to a sparsely filled version of our cemetery, with only very high prestige and very old grave locations occupied (see Figure 6.7).

Thus, if we imagined a version of our cemetery that corresponded more closely to items intended to represent two latent variables, it would only have graves that were relatively pure exemplars of each of the two dimensions, that is, earlier burial and greater prestige. Such an arrangement would result in one group of graves clustered close to the wall of the cathedral (i.e., that of the high-prestige citizens) and another, distinct group clustered at one end of the cemetery perpendicular to the cathedral wall, as shown in Figure 6.7. Locations corresponding simultaneously to greater antiquity and higher prestige (which would be those at the upper right corner of the cemetery as shown in the figures) would be unoccupied. Any one of the remaining graves could clearly be categorized as belonging to one of the clusters, with little relationship to the other cluster.

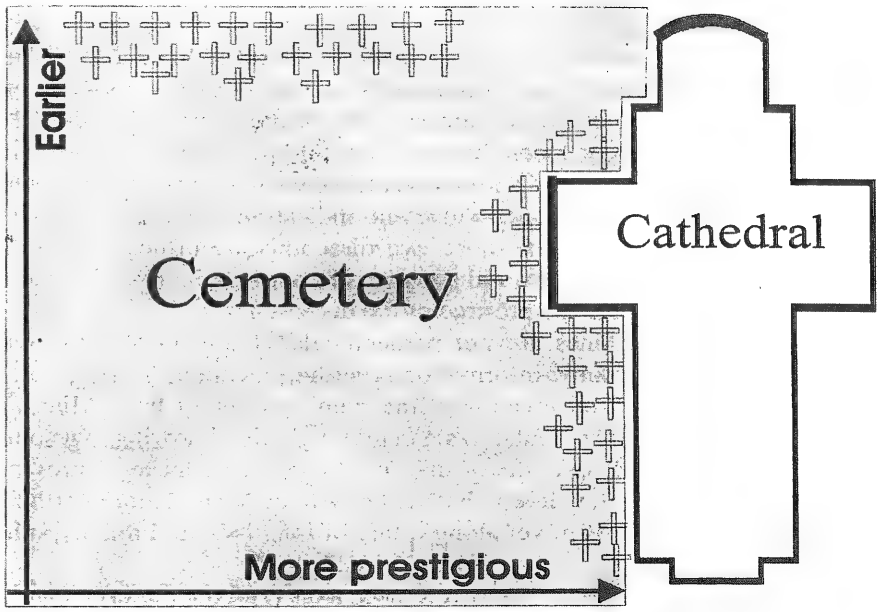

Figure 6.7 Imaginary cemetery with distinct clusters of very old and very prestigious burials

In factor analysis, rotation achieves clarity by seeking factors that result in each item substantially loading on (i.e., correlating with) only one factor. In essence, it tries to find a pattern analogous to the selectively populated graveyard just described. Those items then can be characterized in a meaningful way by noting that they all are relevant to a single factor. There is a hypothetical circumstance in which this attains perfection. Each item has a loading (i.e., correlation) of 1.0 on a single factor and loadings of 0.0 on all other factors. This is called *simple structure*. Rotation algorithms use mathematical criteria that are optimized when the closest possible approximation to simple structure has been obtained. Thus, the end product of a successful factor rotation is an organization of the data that reveals natural groupings of items that share some fundamental characteristic, presumably because of the action of an underlying variable common to those items. At the same time, those items should bear little relationship to any of the other characteristics defining other groups of items.

Orthogonal Versus Oblique Rotation

All the examples thus far have been based on reference lines that are perpendicular to one another. This corresponds to factors that are statistically independent of one another, that is, uncorrelated. Such factors are described as *orthogonal*. Knowledge of location along one line provides no information regarding information along another when the two are perpendicular. Knowing how far north someone is gives no indication of how far west that person is, for example, because those two directions are orthogonal to one another. Similarly, knowing how old a grave is in our imaginary well-ordered cemetery does not indicate the prestige of the deceased.

When two reference lines are not perpendicular, knowing location with respect to one provides some information regarding location on the other. If we replace latitude with an imaginary line running roughly from Miami to Boston (and thus not perpendicular to standard longitude), knowing that someone is at one or the other extreme of that line (or one parallel to it) provides some basis for inferring how likely that person is to be located more northerly or southerly. Directions of travel along a line of longitude and this hypothetical line are correlated.

Factor rotation can likewise allow reference axes (and the factors to which they correspond) to be correlated, and thus not spatially perpendicular. Such a rotation is referred to as *oblique* rather than orthogonal. Oblique rotation may be useful when the underlying latent variables are believed to correlate somewhat with one another. The goal of simple structure requires items that can be meaningfully classified with respect to only a single category. That is, each item should be "about" only one thing and thus load on only one factor. If the variables are somewhat correlated but the factors representing them are forced to be totally independent because of constraints imposed by the factor analytic method, it may be impossible to achieve this goal. That is, more than one factor may be associated with some or all of the items because of the correlation between the factors; we are limited in our ability to approximate simple structure.

If Conscientiousness and Dependability are truly correlated, to return to the earlier co-worker characteristics example, then an item about one is likely to share some variance with the other as well. If, however, the two factors are allowed to be somewhat correlated, the circumstance is roughly analogous to the following reasoning: *Conscientiousness and Dependability are understood to correlate with one another. This fact has been dealt with by allowing the factors to correlate. Now, that aside, to which of these factors is the item in question most strongly related?* Thus, allowing the factors themselves to be correlated with one another makes it possible for items to be less ambiguously

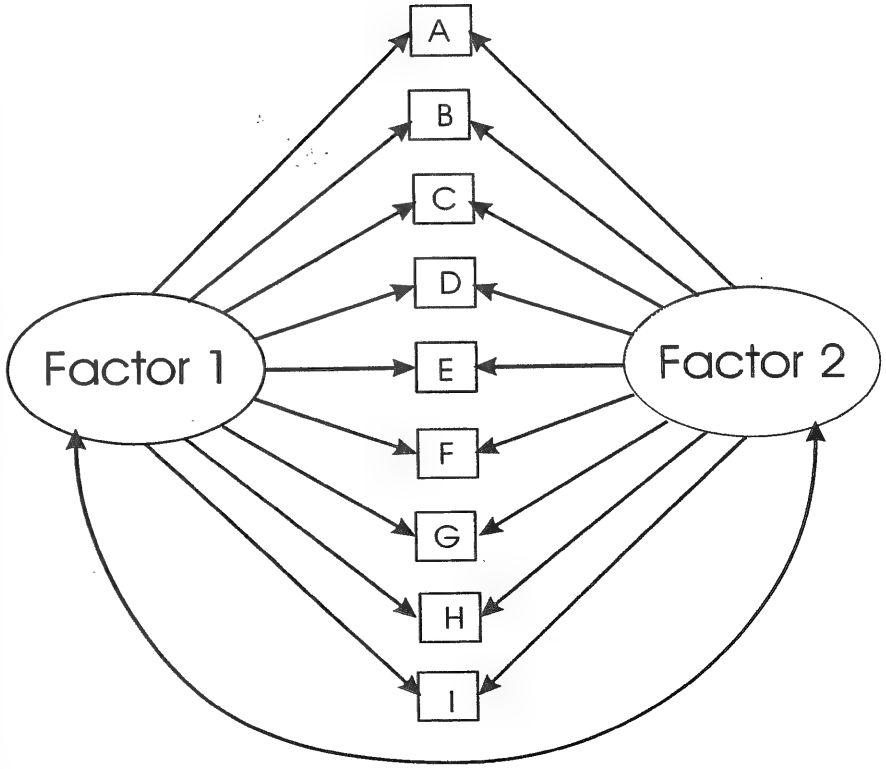

Figure 6.8 Two-factor model with factors allowed to correlate

identified with one or the other factor, thus bringing us closer to our goal of simple structure. The correlation between a given pair of items is accommodated at the factor level, and, as a result, the item need not be associated with both factors.

What is lost when factors are rotated obliquely is the elegance and simplicity of uncorrelated dimensions. One very nice feature of uncorrelated factors is that their combined effects are the simple sum of their separate effects. The amount of information in a specific item's value that one factor explains can be added to the information that another factor explains, to arrive at the total amount of information explained by the two factors jointly. With oblique factors, this is not the case. Because they are correlated, there is redundancy in the information contained in the two factors. For an item related to both

Conscientiousness and Dependability, the amount of variation explained by those two factors together is less than the sum of the parts. Some, perhaps a great deal, of the information accounted for by one factor overlaps with the information accounted for by the other. A simple sum would include the overlapping information twice, which would not accurately reflect the total influence of the factors on that item.

Another complication of correlated factors is the added complexity of the causal relationships between items and factors. When the factors are independent, the only relationship between a factor and a specific item is direct. Changes in the level of the factor will result in changes in the item along a single, direct causal pathway. When factors are correlated, however, this is not the case. If two hypothetical factors both influence Item A, for example, and the factors are correlated, each factor exerts an indirect as well as a direct influence on A. That is, Factor 1 can influence Factor 2 and, through Factor 2, indirectly affect Item A. This is in addition to the direct effect of Factor 1 on that item.

Of course, by a parallel process, Factor B also can affect the item not only directly but also indirectly through its relationship to Factor A. The same sort of direct-plus-indirect influence of factors also may apply to all other items. As a result, speaking of the relationship between an item and a factor usually has to be qualified explicitly to either include or exclude such indirect effects. Otherwise, there is ambiguity and thus potential confusion.

Choosing a Type of Rotation

As a practical matter, the choice between orthogonal and oblique rotation should depend on one or more considerations. Among these is how one views the concepts the factors represent. If theory strongly suggests correlated concepts, it probably makes sense for the factor analytic approach (specifically, rotation) to follow suit. Thus, if we are analyzing items related to Conscientiousness and Dependability, allowing factors to correlate would best fit our sense of what these concepts imply. Alternatively, theory might suggest orthogonal factors. Dependability and Fun, for example, may be more independent and thus might warrant an orthogonal solution. When theory does not provide strong guidance, as when the scale under development represents concepts not previously studied, the magnitude of the correlations between factors may serve as a guide. Specifically, an oblique rotation can be specified and the resultant correlations among factors examined. If these are quite small (e.g., less than .15), the data analyst might opt for orthogonal rotation. This slightly compromises the approximation of simple structure, but it results in a simpler model. For example, some items may show secondary loadings

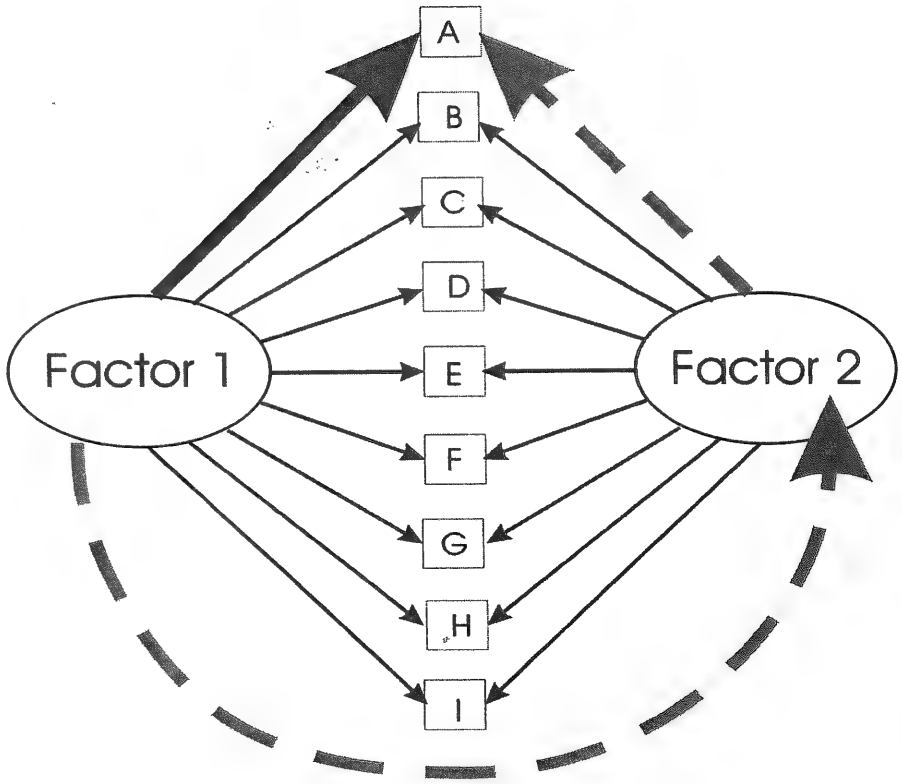

Figure 6.9 Because of the correlations between the factors, Factor 1 influences Item A both directly (the dark, solid pathway) and indirectly (the lighter, dashed pathway)

(i.e., loadings on a factor other than the one on which they load most strongly) that are slightly increased relative to the oblique solution but still small enough to associate each item unambiguously with only one factor. Thus, loadings of a particular item on three factors rotated to an oblique solution might be .78, .16, and .05. When an orthogonal solution is chosen, the loadings might be .77, .19, and .11. Although the second pattern departs slightly more than the first from simple structure, the item in question can still be linked unambiguously to the first factor. Thus, little has been sacrificed in opting for the simpler (i.e., orthogonal) model in this case. If the factors are more substantially correlated,

opting for the oblique solution might yield a substantial improvement in approximation to simple structure. For example, a secondary loading of .40 obtained with an orthogonal rotation might diminish to .15 with an oblique solution. This will not always be the case, however, and only examining the difference between the two rotation methods can definitively indicate how much they differ in approximating simple structure.

A final practical issue concerns the magnitude of the correlation between two factors and how large it needs to be before combining them into one larger factor. There is no simple answer to this question because the relationships of items to factors also needs to be considered. Under some circumstances, however, an oblique rotation may reveal that even when two factors are highly correlated, some items have substantial loadings on both. In that case, it might make good sense to extract one factor to see if the two that were highly correlated factors merge into one. It might very well be, for example, that real-world data would support a single factor combining items about Conscientiousness and Dependability rather than separating them.

INTERPRETING FACTORS

In the example involving Conscientiousness and Dependability items, we have assumed that we knew a priori exactly what the latent variables were. Often, this is not the case, and then we will rely on the factor analysis to give us clues regarding the nature of those latent variables. This is done by examining the items that most strongly exemplify each factor, that is, that have the largest loadings on a particular factor. The items with the highest loadings are the ones that are most similar to the latent variable (and, thus correlate most strongly). Therefore, they can provide a window into the nature of the factor in question. This is most easily done when there are several items that clearly tap one common variable with quite substantial loadings (e.g., greater than .65) on the same factor. Returning to the example of characteristics considered important in a co-worker, if "smart," "has a mind like a steel trap," "is well educated," and perhaps one or two other items related to intellectual capacity all loaded substantially on the same factor, with no other items having large loadings on that factor, it would be fairly easy to conclude that "Importance Ascribed to Intellect" or some equivalent description was an apt label for that factor.

Although choosing a label for a factor may seem straightforward in some cases, assigning a name is *not* the same as establishing validity. Whether the item set continues to perform as the assigned name implies will ultimately

determine validity. When factors explain relatively little variance and have multiple, seemingly disparate items that load comparably on them, the factor analyst should be especially cautious in interpretation. If the analysis yielded one factor with items that seem dissimilar, it probably is best not to take the factor too seriously as an indicator of a latent variable.

Another point worth remembering at the interpretation stage is that a factor analysis can only find the structure accounting for associations among the items analyzed—it will not necessarily reveal the nature of phenomena per se. A researcher attempting to determine the fundamental dimensions of personality, for example, could not obtain an Extroversion factor if no items pertaining to extroversion were included.

Sometimes, inclusion of a specific phrase can create a false appearance of a conceptually meaningful factor. When some statements are worded in the first person and others are not, for example, that may account for the pattern of associations observed. As an illustration, consider the following hypothetical items:

1. I like apples.
2. Oranges taste good.
3. I prefer apples to some other types of fruit.
4. There are many people who like oranges.
5. I enjoy an apple every now and then.
6. Oranges generally have a pleasant fragrance.
7. I find the crispness of apples appealing.
8. A fresh orange can be a tasty treat.

Were the odd items to load on one factor and the even items on a second factor, we would not know if the "I" wording of the odd items was the cause of the two factors or if people were expressing differential attitudes toward the two types of fruit mentioned. Both explanations are plausible but confounded. This is a case where we may or may not be comparing apples to oranges.

PRINCIPAL COMPONENTS VERSUS COMMON FACTORS

There are two broad classes of data analytic techniques that some authors regard as fundamentally the same but others view as fundamentally different. These are factor analysis and principal components analysis. The term *factor analysis* is used sometimes to embrace both techniques and at other times to

describe one as opposed to the other. The terms *common factors* and *components* are often used as a less ambiguous way of referring specifically to the composites arising from factor analysis and principal components analysis, respectively. There is a basis for asserting both the similarity and the dissimilarity of these methods.

Principal components analysis (PCA) yields one or more composite variables that capture much of the information originally contained in a larger set of items. The components, moreover, are defined as weighted sums of the original items. That is, principal components are linear transformations of the original variables. They are grounded in actual data and are derived from the actual items. They are merely a reorganization of the information in the actual items.

Common factor analysis also yields one or more composite variables that capture much of the information originally contained in a larger set of items. However, these composites represent hypothetical variables. Because they are hypothetical, all we can obtain are estimates of these variables. A common factor is an idealized, imaginary construct that presumably causes the items to be answered as they are—the nature of the construct is inferred by examining how it affects certain items.

The Similarities and Differences of Components and Factors

The above descriptions highlight some differences between components and factors. One of those differences is that factors represent idealized, hypothetical variables we estimate, whereas components are alternative forms of the original items, with their information combined. The idea behind extracting common factors is that we can remove variance from each item that is not shared with any of the other items. From the perspective of factor analysis, as was the case with reliability, unshared variation is essentially error. Thus, the combinations we arrive at in extracting common factors are estimates of hypothetical error-free underlying variables. It is in this sense that common factors are idealized—they are estimates of what an error-free variable determining a set of items might look like. Furthermore, factors *determine* how items are answered, whereas components are *defined by* how items are answered. Thus, in PCA, the components are end products of the items and the actual scores obtained on items determine the nature of the components. In common factor analysis, however, we invoke the concept of an idealized hypothetical variable that is the cause of the item scores. A factor is an estimate of that hypothetical variable and represents a cause, not an effect, of item scores.

What about the similarities? There are several. First, the computational difference between the two is very minor. Remember, in common factor

TABLE 6.1
Correlation Matrixes for Principal Components Analysis and Common Factor Analysis

1.0	.70	.83	.48	.65
.70	1.0	.65	.33	.18
.83	.65	1.0	.26	.23
.48	.33	.26	1.0	.30
.65	.18	.23	.30	1.0

.45	.70	.83	.48	.65
.70	.52	.65	.33	.18
.83	.65	.62	.26	.23
.48	.33	.26	.48	.30
.65	.18	.23	.30	.58

NOTE: The correlation matrix on the left, which is used for principal components analysis, retains unities in the main diagonal. The correlation matrix on the right, which is used for common factor analysis, has communality estimates, rather than unities, along the main diagonal.

analysis, the goal is to estimate an idealized, error-free variable. But we must generate this estimate from the actual data. As we have noted, factor analytic methods generally are based on a correlation matrix representing all the associations among the items to be factored. Back in Chapter 3, I pointed out that all of the off-diagonal values in a covariance or correlation matrix represent only shared or communal variance. As I pointed out then, a correlation matrix is simply a standardized version of a variance-covariance matrix. The correlations themselves are standardized covariances and the unities are standardized item variances. Each standardized item variance represents all the variability, both shared and unique, that an item manifests. To create an idealized, error-free variable, the unique portion of variance contained in the item variances along the main diagonal of the correlation matrix must be purged. More specifically, each unity must be replaced by a *communality estimate*, a value less than 1.0 that approximates only a given variable's shared variation with other variables included in the factor analysis. For example, if we estimated that a particular variable shared 45% of its total variation with other items in the correlation matrix, we would assign it a communality estimate of .45 and place that value in the matrix, replacing the 1.0 that represented the item's total variance. We would do this for every variable, replacing each unity with a communality estimate. (Often, communality estimates are obtained by regressing the variable in question on all the remaining variables so as to obtain the squared multiple correlation, R^2, which serves as the estimate.) This purging process creates an altered correlation matrix that is used for extracting common factors rather than components, as shown in Table 6.1.

Substituting communality estimates for unities is the only computational difference separating the extraction of common factors from the extraction of principal components.

What about the "cause versus effect" issue? Is it or is it not the case that we obtain both factors and components by analyzing scores on observed items? It is the case. As the explanation of communality estimates demonstrates, empirical relationships among the items ultimately are the foundation for common factors. This, of course, is also true for components. So, computationally, both are grounded in empirical data. Furthermore, most data analysts conceptualize both components and common factors as ways of understanding the variables underlying a set of items. That is, both components and factors are customarily thought of as revealing the cause for observed scores on a set of items. Components analysis and factor analysis, in fact, are often used interchangeably; under most circumstances in which items have something meaningful in common, the different methods support the same conclusions. So, while there are both technical similarities and differences between the two, the distinctions between them are often overlooked, with little if any adverse consequences.

One important difference between them, however, is the nature of the variance explained by components versus factors. The former account for a specified portion of the *total* variance among the original variables, whereas the latter account for the *shared* or *common* variance. Reducing the diagonal values of the correlation matrix, as one does when extracting common factors, reduces both the numerator and the denominator of the ratio expressing proportion of variance. But it reduces the denominator to a greater degree because of the specific calculations involved in computing the relevant variances. As a result, the "proportion of variance explained" by a set of comparable components and factors will not be equal or conceptually equivalent. Factors will explain a larger proportion of a more restricted variance (i.e., shared variance), while components will explain a smaller proportion of total variance. When discussing factor analytic results and reporting the proportion of variance explained by the factors, it is critical to be clear about what type of analysis (components or common factors) and, thus, what type of variance (shared or total) is being explained.

Another difference between the two types of analysis that is worth noting is that, in some statistical packages, some of the output obtained from the extraction of common factors, but not components, will appear nonsensical. In both types of analysis, the cumulative amount of variance explained will mount as each successive factor or component is extracted. With common factors, this proportion often exceeds 1.0 at some point, continues to rise as successive factors are considered, and then, as if by magic, returns to a value of precisely 1.0 as the *k*th (i.e., last possible) factor is extracted. Although this looks

strange, it is simply an artifact of the computation method and can be ignored. If the data analyst has used reasonable criteria for deciding how many factors to extract, the number chosen will typically precede the point in the extraction sequence where this anomaly arises. It is possible, however, for the selected number of factors to explain virtually all (i.e., 100%) of the shared variance among the original items.

CONFIRMATORY FACTOR ANALYSIS

Another bifurcation of factor analytic methods differentiates between *exploratory* and *confirmatory* techniques. These terms originally referred to the intent of the data analyst rather than the computational method. Thus, the same analysis might be used on the same set of items either to determine what their underlying structure is (exploratory) or to confirm a particular pattern of relationships predicted on the basis of theory or previous analytic results (confirmatory). With increasing frequency, these terms are now used to differentiate between different types of analytic tools rather than different research objectives. When people use the term *confirmatory factor analysis*, they are often talking about methods based on structural equation modeling (SEM). Although these methods should be used in a confirmatory rather than exploratory fashion, standard factor analytic techniques can be used for either. Thus, "confirmatory" does not necessarily equate with "SEM based."

The SEM-based methods, however, offer some real benefits over traditional factor analytic methods in certain situations. These benefits arise because the SEM models are extremely flexible. Conditions that are assumed by traditional factoring methods, such as independence of the item error terms from one another, can be selectively altered in SEM-based approaches. Also, traditional methods for the most part constrain the data analyst to either allow factors to correlate with one another or require that they are all independent of one another. SEM-based approaches can mix correlated and uncorrelated factors if theory indicates such a model applies.

As noted earlier, SEM-based approaches also can provide a statistical criterion for evaluating how well the real data fit the specified model. Used judiciously, this can be an asset. Sometimes, however, it can lead to overfactoring. Extracting more factors often improves a model's fit. Applying a strictly statistical criterion may obscure the fact that some statistically significant factors may account for uninterestingly small proportions of variance. Especially in the early stages of instrument development, this may be contrary to the goals of the investigator who is concerned with finding the fewest number of

most information-laden factors rather than accounting for the maximum amount of variance possible.

Another double-edged sword in SEM-based approaches is the common practice of testing alternative models and comparing how they fit the data. Again, used prudently, this can be a valuable tool. Conversely, used carelessly, it can result in model specifications that may make little theoretical sense but result in statistically better model fit. As examples, removing the constraint that item errors be uncorrelated with one another might yield quite small values for the correlations, but the model might still statistically outperform a constrained model. One researcher may decide to ignore the small correlations in favor of a simpler model, while another is persuaded by a statistical criterion to reject the more parsimonious alternative. As another example, a model that separates two distinct but highly correlated factors (perhaps like Conscientiousness and Dependability) might fit better than one that combines the two. If the correlation between them is very high, the decision to keep them separate may seem arbitrary. For example, a correlation of, say, .85 between two indicators of the same construct would usually be considered as good evidence of their equivalence. But a model that specified separate factors that correlated with each other at .85 might fit the data better than a model that combined the two into a single factor.

These comments are not intended to suggest that SEM-based approaches to confirmatory factor analysis are bad. The advent of these methods has made enormous contributions to understanding a variety of measurement issues. I am suggesting, however, that the inherent flexibility of these approaches creates more opportunities for making poor decisions, particularly when the data analyst is not very familiar with these methods. With the possible exception of principal components analysis (in which factors are linear combinations of the items), no factoring method produces a uniquely correct solution. These methods merely produce plausible solutions, of which there may be very many. There is no guarantee that a more complex model that statistically outperforms a simpler alternative is a more accurate reflection of reality. It may or it may not be. With all factor analytic approaches, common sense is needed to make the best decisions. The analyses are merely guides to the decision-making process and evidence in support of the decisions made. They should not, in my judgment, entirely replace investigator decision making. Also, it is important that the bases for decisions, statistical or otherwise, are described accurately in published reports of confirmatory factor analysis.

One last note on this subject: Researchers in some areas of inquiry (e.g., personality research) consider obtaining consistent results from traditional factoring methods as stronger confirmatory evidence than demonstrating good model fit according to a statistical criterion. For example, Saucier and Goldberg (1996) state that "because exploratory factor analysis provides a more rigorous replication test than confirmatory analysis, the former technique may often be preferred to the latter" (p. 35). The reasoning is that if data from different samples of individuals on different occasions produce essentially identical factor analytic results using exploratory approaches, the likelihood of those results being a recurring quirk is quite small. Remember that in SEM-based approaches to this same situation, the data analyst specifies the anticipated relationships among variables and the computer program determines if such a model can be reconciled with the empirical data. In other words, the computer is given a heavy hint as to how things should turn out. In contrast, rediscovering a prior factor structure without recourse to such hints, as may happen with repeated exploratory analyses, can be very persuasive.

USING FACTOR ANALYSIS IN SCALE DEVELOPMENT

The following example should make some of the concepts discussed in this chapter more concrete. Some colleagues and I (DeVellis, DeVellis, Blanchard, Klotz, Luchok, & Voyce, 1993) developed a scale assessing parents' beliefs about who or what influences their children's health. Although the full scale has 30 items and assesses several aspects of these beliefs, for this presentation I will discuss only 12 of the items:

A. I have the ability to influence my child's well-being.
B. Whether my child avoids injury is just a matter of luck.
C. Luck plays a big part in determining how healthy my child is.
D. I can do a lot to prevent my child from getting hurt.
E. I can do a lot to prevent my child from getting sick.
F. Whether my child avoids sickness is just a matter of luck.
G. The things I do at home with my child are an important part of my child's well-being.
H. My child's safety depends on me.
I. I can do a lot to help my child stay well.
J. My child's good health is largely a matter of good fortune.
K. I can do a lot to help my child be strong and healthy.
L. Whether my child stays healthy or gets sick is just a matter of fate.

These were administered to 396 parents and the resultant data were factor analyzed. The first objective of the factor analysis was to determine how many factors were underlying the items. SAS was used to perform the factor

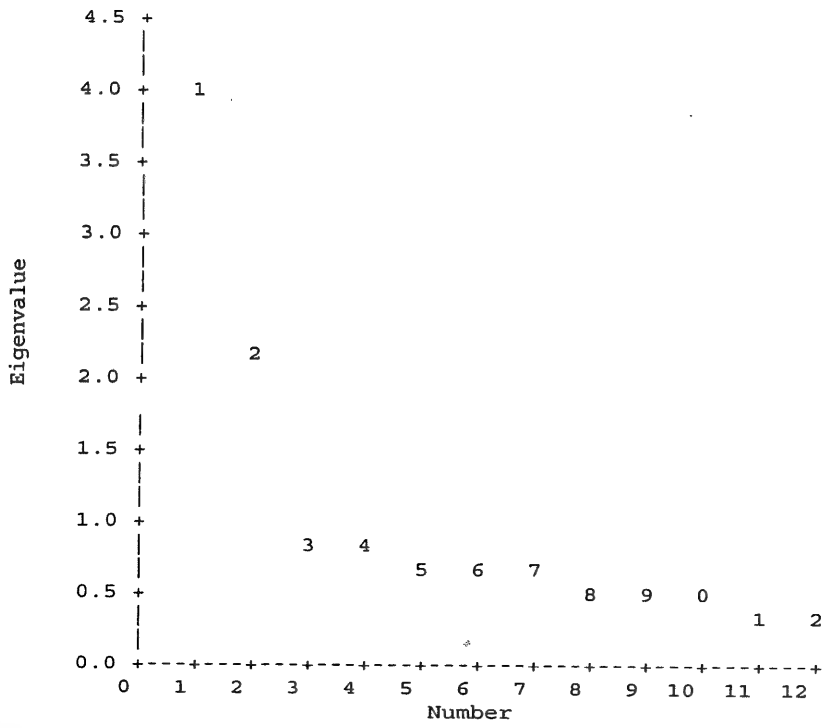

Figure 6.10 A scree plot from factor analysis of selected items

analysis, and a scree plot was requested. A scree plot similar to the type printed by SAS appears in Figure 6.10. Note that 12 factors (i.e., as many as the number of items) are plotted. However, 2 of those factors are located on the initial portion of the plot and the remainder form the scree running along its bottom. This strongly suggests that 2 factors account for much of the variation among the items.

Having determined how many factors to retain, we reran the program specifying 2 factors and *varimax* (orthogonal) rotation. Had we failed to approximate simple structure, we might have performed an oblique rotation to improve the fit between items and factors. However, in this case, the simpler orthogonal rotation yielded meaningful item groupings and strong, unambiguous loadings.

This is evident from the following table of factor loadings, in which each row contains the loadings of a given item on the two factors. An option

TABLE 6.2
Item Loadings on Two Factors

	Rotated Factor Pattern	
	Factor 1	*Factor 2*
Item I	**0.78612**	−0.22093
Item K	**0.74807**	−0.18546
Item D	**0.71880**	−0.02282
Item E	**0.65897**	−0.15802
Item G	**0.65814**	0.01909
Item A	**0.59749**	−0.14053
Item H	**0.51857**	−0.07419
Item F	−0.09218	**0.82181**
Item J	−0.10873	**0.78587**
Item C	−0.07773	**0.75370**
Item L	−0.17298	**0.73783**
Item B	−0.11609	**0.63583**

available in SAS has reordered the items in the table so that those with high loadings on each factor are grouped together.

In Table 6.2, factor loadings greater than 0.50 have been bolded. Each factor is defined by the items that load most heavily on it (i.e., those bolded). By referring to the content of those items, one can discern the nature of the latent variable that each factor represents. In this case, all of the items loading strongly on Factor 1 concern the parent as an influence over whether a child remains safe and healthy. Those loading primarily on Factor 2, on the other hand, concern the influence of luck or fate on the child's health.

These two homogeneous item sets can be examined further. For example, alpha could be computed for each grouping. Computing alpha on these item groupings using SAS yields the results in Table 6.3.

Both scales have acceptable ålpha reliability coefficients. Note that the SAS Corr procedure calculates alpha for unstandardized and standardized items. The latter calculation is equivalent to using the correlation-based alpha formula. For both scales, these two methods of computing alpha yield quite similar values. Note also that for neither scale would alpha increase by dropping any item. Alphas nearly as high as those obtained for the full scales result from dropping one item, that is, item H from scale 1 and item B from scale 2. However, retaining these items provides a bit of additional insurance that the reliability will not drop below acceptable levels on a new sample and does not increase the scales' length substantially.

TABLE 6.3
Coefficient Alphas for All Items and for All $k - 1$ Combinations of Items, for Two Different Sets of Items

Cronbach Coefficient Alpha
For Raw Variables: 0.796472 For Standardized Variables: 0.802006

	Raw Variables		*Standardized Variables*	
Deleted Variable	*Correlation with Total*	*Alpha*	*Correlation with Total*	*Alpha*
Item I	0.675583	0.741489	0.676138	0.749666
Item K	0.646645	0.748916	0.644648	0.755695
Item E	0.545751	0.770329	0.535924	0.775939
Item D	0.562833	0.763252	0.572530	0.769222
Item G	0.466433	0.782509	0.474390	0.787007
Item H	0.409650	0.793925	0.404512	0.799245
Item A	0.437088	0.785718	0.440404	0.793003

For Raw Variables: 0.811162 For Standardized Variables: 0.811781

	Raw Variables		*Standardized Variables*	
Deleted Variable	*Correlation With Total*	*Alpha*	*Correlation With Total*	*Alpha*
Item F	0.684085	0.748385	0.682663	0.749534
Item C	0.596210	0.775578	0.594180	0.776819
Item J	0.636829	0.762590	0.639360	0.763036
Item L	0.593667	0.776669	0.592234	0.777405
Item B	0.491460	0.806544	0.493448	0.806449

All of the caveats applicable to scales in general at this point in their development are applicable to factor analytically derived scales. For example, it is very important to replicate the scales' reliability using an independent sample. In fact, it probably would be useful to replicate the whole factor analytic process on an independent sample to demonstrate that the results obtained were not a one-time chance occurrence.

SAMPLE SIZE

The likelihood of a factor structure replicating is at least partially a function of the sample size used in the original analysis. In general, the factor pattern that

emerges from a large-sample factor analysis will be more stable than that emerging from a smaller sample. Inevitably, the question arises, "How large is large enough?" This is difficult to answer (see, e.g., MacCallum, Widaman, Zhang, & Hong, 1999). As with many other statistical procedures, both the relative (i.e., to the number of variables analyzed) and the absolute number of subjects should be considered, but factors such as item communalities also play a role (MacCallum et al., 1999). The larger the number of items to be factored and the larger the number of factors anticipated, the more subjects should be included in the analysis. It is tempting, based on this fact, to seek a standard ratio of subjects to items. However, as the sample gets progressively larger, the ratio of subjects to items often can diminish. For a 20-item factor analysis, 100 subjects would probably be too few, but for a 90-item factor analysis, 400 might be adequate. Tinsley and Tinsley (1987) suggest a ratio of about 5 to 10 subjects per item up to about 300 subjects. They suggest that when the sample is as large as 300, the ratio can be relaxed. In the same paper, they cite another set of guidelines, attributed to Comrey (1973), that classifies a sample of 100 as poor, 200 as fair, 300 as good, 500 as very good, and 1,000 as excellent. Comrey (1988) stated that a sample size of 200 is adequate in most cases of ordinary factor analysis that involve no more than 40 items. Although the relationship of sample size to the validity of factor analytic solutions is more complex than these rules of thumb indicate, they will probably serve investigators well in most circumstances.

It is certainly not uncommon to see factor analyses used in scale development based on more modest samples (e.g., 150 subjects). However, the point is well taken that larger samples increase the generalizability of the conclusions reached by means of factor analysis. Of course, replicating a factor analytic solution on a separate sample may be the best means of demonstrating its generalizability.

CONCLUSION

Factor analysis is an essential tool in scale development. It allows the data analyst to determine the number of factors underlying a set of items so that procedures such as computing Cronbach's alpha can be performed correctly. In addition, it can provide us with insights into the nature of the latent variables underlying our items.

7

An Overview of Item Response Theory

Item response theory (IRT) is an alternative to classical measurement theory (CMT), which is also called classical test theory (CTT). IRT has received increasing attention in recent years (see, e.g., Hambleton, Swaminathan, & Rogers, 1991; Embretson & Reise, 2000). The basic idea underlying classical measurement theory is that an observed score is the result of the respondent's true score plus error. That error is not differentiated into subcategories, such as differences across time, settings, or items. Instead, all sources of error are collected in a single error term. IRT methods differentiate error more finely, particularly with respect to item characteristics.

Although IRT has been used primarily for ability measures (such as the Scholastic Aptitude Tests), nothing about the theory precludes its application to other domains. Whereas classical measurement theory concerns itself primarily with composites and, more specifically, scales, IRT focuses primarily on individual items and their characteristics. In CMT, items are a means to an end, in a sense. That is, they are roughly equivalent indicators of the same underlying phenomenon that gain strength through their aggregation as a scale. A scale's reliability is increased by redundancy. In IRT, each item's relationship to the variable of interest (often called the *attribute*) is assessed. Reliability is enhanced not by redundancy but by identifying better items. More IRT-based items typically increase the number of points along an attribute's continuum that can be differentiated, but they do not increase reliability in the way we considered it earlier. As an example, adding more difficult questions to a mathematics test extends the test's useful range upward but does not necessarily have any effect on internal consistency. A requirement of scale development under CMT is that the items share a common cause and thus are correlated with one another. Moreover, a given scale should have only a single underlying dimension. IRT shares this characteristic; items that will be grouped together must share a single latent variable. CMT items, therefore, are redundant, and indeed that redundancy is an important part of a scale's reliability. However, whereas each CMT item is designed to be very similar to every other item and to tap the underlying variable in the same way, IRT items are designed to tap different degrees or levels of the attribute. Precision is enhanced not by having redundant items that are summed but by having

nonredundant items with specific, demonstrable characteristics. We will explore these characteristics a bit later on in this chapter.

Because item response theory originated in the context of ability testing, its vocabulary contains terms usually associated with that content area. Also, because items on ability tests are usually graded as correct or incorrect (even though their original format may involve more than two response options), the classic applications and examples of IRT involve items that take on one of two states (e.g., "pass" or "fail"). It is easiest to discuss IRT for items of this type, although there is no reason why the methods stemming from the theory should not be extended (as, indeed, they have) to items with other response formats (e.g., Likert scales) tapping other content domains.

The goal of IRT is to enable a researcher to establish certain characteristics of items independent of who completes them. This is analogous to physical measurement in which one can assess an attribute of an object (e.g., length or weight) without regard to the specific nature of the object. Twenty pounds, for example, means the same thing no matter what is being weighed. Thus, a conventional bathroom scale yields information about a specific characteristic of objects (i.e., weight) irrespective of the nature of the object being weighed. IRT aspires to do the same thing with questionnaire items.

IRT is really a family of models rather than a theory specifying a single set of procedures. One important way in which the alternative IRT models differ is in the number of item parameters with which they are concerned. A common approach in recent years has been the three-parameter model that, not surprisingly, concentrates on three aspects of an item's performance. These are the item's *difficulty*, its *capacity to discriminate*, and its *susceptibility to false positives*. An early, and still popular, member of the IRT family is Rasch modeling (Rasch, 1960; Wright, 1999), which quantifies only the difficulty parameter.

ITEM DIFFICULTY

Although the term is a clear carryover from ability testing, the concepts it represents are more widely applicable. *Item difficulty* refers to the level of the attribute being measured that is associated with a transition from "failing" to "passing" that item. Most of us have seen old movies depicting carnivals or amusement parks featuring a certain feat of strength. The "measurement device" is a vertical track along which a weight travels. At the top of the track is a bell. Initially, the weight rests at the bottom of the track, on the end of a

plank serving as a kind of seesaw. "Respondents" strike the end of this seesaw opposite the weight with a large mallet, thus sending the weight flying upward along the track. Their goal is to propel the weight with enough force to strike the bell, ringing it. For our purposes, we can think of the entire apparatus as the "item."

The *difficulty* of the item is the amount of strength the "respondent" must possess (or more accurately, the force she or he must transmit) in order to "pass" the item (i.e., ring the bell). Clearly, one could construct different items with different degrees of difficulty (e.g., more difficult "items" with longer tracks or heavier weights). It should be possible, however, to calibrate the difficulty of a particular apparatus that is independent of any characteristic of the person who happens to be swinging the mallet at the moment.

Because this "item" is a physical object, it would be fairly easy to determine with reasonable precision how much force was needed to cause the bell to ring (ignoring, for the moment, the effect of striking the seesaw in slightly different locations relative to the fulcrum). So, the carnival operator could presumably order a 10-pound apparatus or a 100-pound apparatus to achieve either a high or a low "passing rate" among people playing the game. Each might be specifically suitable for different groups of customers, for example, children attending a school fair or adults attending an athletic camp.

One can characterize a questionnaire item in a similar way. Imagine, for example, an item measuring depression. One could construct the item to be relatively "easy" or relatively "difficult." In the first instance, only a modest amount of the attribute "depression" would be needed to "pass" the item (which might be operationalized as the respondent's indicating having experienced a particular feeling at least once a week). For example, an item such as "I feel discouraged by all I have to do" would probably be an "easy" item in this sense. But would not the likelihood of the person having had that feeling "one or more times a week" depend on who was asked? If, for example, we posed the question to people who were clinically depressed, we would probably find a larger proportion of that sample "passing" the item than if we administered it to the general population. The goal of determining item difficulty is to establish in an absolute sense how much of the attribute is required to pass the item. If this can be done, then a person's passing the item has a constant meaning with respect to level of depression, irrespective of who that person is or the average level of depression in the sample under study. In other words, the person can be characterized not merely in reference to a specific sample but in terms of a metric independent of any specific sample.

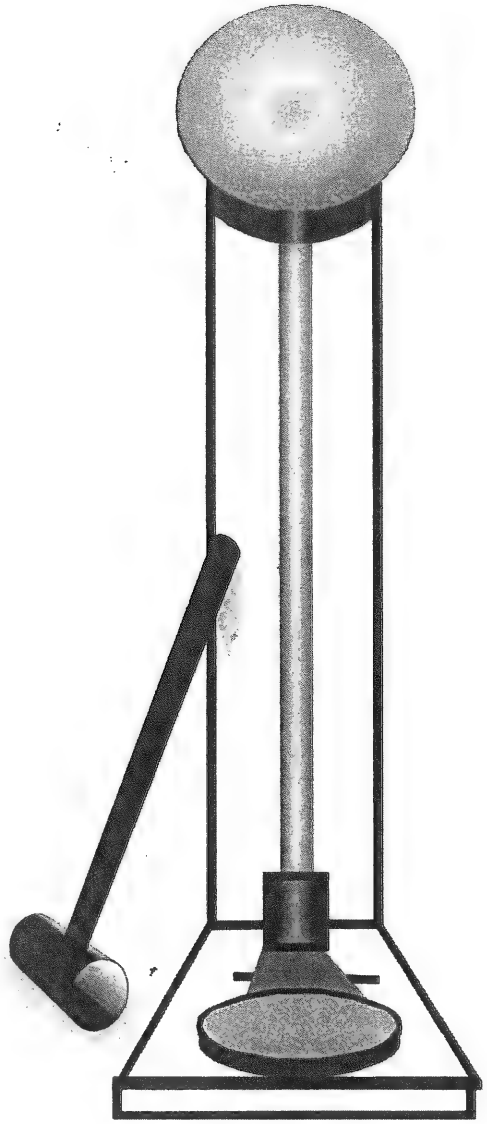

Figure 7.1 A hypothetical apparatus for testing strength, in which striking the pad with the hammer with sufficient force causes the bell to ring

ITEM DISCRIMINATION

The second parameter IRT addresses is the degree to which an item unambiguously classifies a response as a "pass" or "fail." Stated differently, the less the ambiguity about whether a person is truly a pass or fail, the higher the discrimination of the item in question. Using our carnival bell analogy, there may be occasions when the weight barely makes contact with the bell and so, causing observers to disagree as to whether the bell actually rang. Some might hear a faint ring while others hear nothing. Within the range of force that propels the weight so that it touches the bell but does not produce what all agree is a clear ringing sound, the device is providing ambiguous information. Looking at this ambiguity another way, the same force applied many times might result in observers determining that the bell rang on some occasions but did not ring on others. A somewhat greater force will consistently produce an unambiguous ring, whereas a somewhat weaker force will produce an equally unambiguous failure to ring. But there is a small range of force in which the device is ambiguous. An alternative device might operate differently and produce less ambiguous results. For example, the weight touching the bell might close an electrical circuit that trips a relay, causing a light to luminate and remain lit until it is reset. If well engineered, such a device would probably yield consistent results over a fairly small range of forces and thus would discriminate better than the standard device. Alternatively, a device that had no bell at all but instead required the observers to raise their hands if the weight crossed a predetermined line marked next to its track would probably produce more ambiguity and thus would discriminate less well. So, a device or item that discriminates well has a very narrow portion of the range of the phenomenon of interest in which the results are ambiguous. A less discriminating device or item has a larger region of ambiguity.

FALSE POSITIVES

The third parameter in IRT is false positives. A *false positive* is a response indicating that some characteristic or degree of an attribute exists when in actuality it does not. Here, we need a new carnival analogy. You may have seen booths in which a person sits behind a protective sheet of plexiglass, above a tank of water, on a platform connected to a lever extending to one side on which a target is painted.

Contestants are invited to throw baseballs at the target, which, if hit, causes the platform to collapse and the person to fall into the tank of water below him

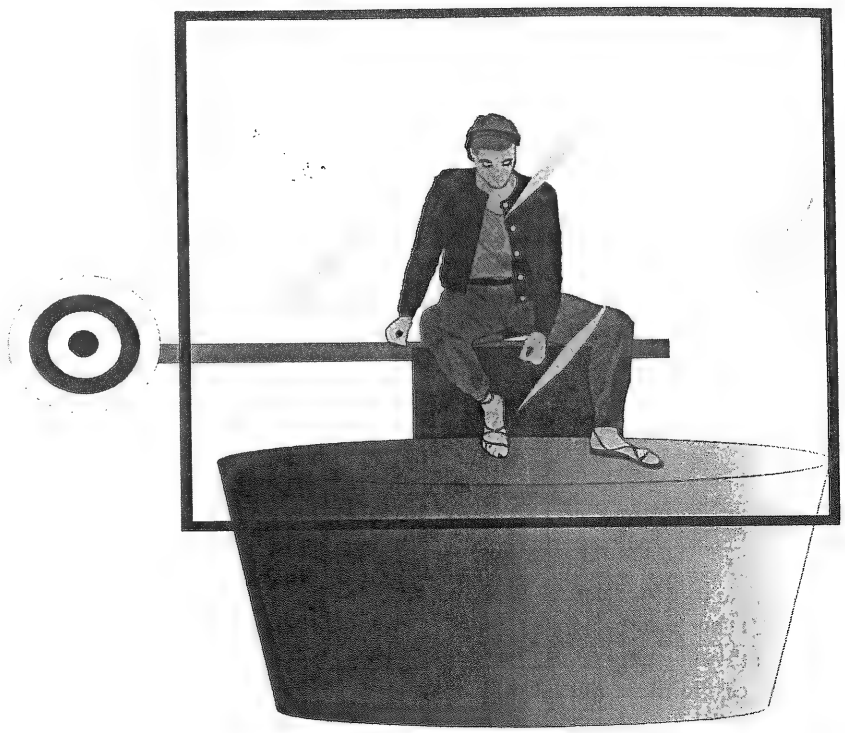

Figure 7.2 A hypothetical apparatus for measuring throwing accuracy, in which striking the target with a ball causes the platform to collapse and the person sitting on it to fall into a tank of water

or her. We can think of this device as an "item" that measures throwing accuracy. Causing the person on the platform to fall into the tank of water constitutes a "pass" on this "item." (By now, you should be able to describe how variations in the apparatus could increase or decrease the *difficulty* and *discrimination* of the device.) With this particular device, one can imagine how "false positives" might occur, that is, how a respondent with virtually no ability could score a "pass" by causing the person perched above the tank to be dunked. One way might be that the "respondent" throws wildly but the ball just happens to hit the target (it has to go somewhere, after all). Or, alternatively, the apparatus might malfunction and the platform might collapse spontaneously. In these cases, the player/respondent would "pass" based not on

ability but on some unrelated circumstance. Thus, it is possible to "pass" this test of throwing accuracy even if one has little or no ability. In the context of ability testing, false positives most commonly occur as the result of successfully guessing the correct response to a question despite not really knowing the answer. (In measurement contexts where the opportunities for guessing or other types of false positives are minimal, such as using scales to measure weight, a two-parameter model is often sufficient.)

Each of these three item parameters—difficulty, discrimination, and false positives—bears a fairly obvious relationship to measurement error. If (a) the difficulty of an item is inappropriate, (b) the area of ambiguity between a pass and a fail is large, or (c) the item indicates the presence of a characteristic even when it is absent, then the item is prone to error. IRT quantifies these three aspects of an item's performance and thus provides a means for selecting items that will likely perform well in a given context.

ITEM CHARACTERISTIC CURVES

This quantification is then summarized in the form of an *item characteristic curve* (ICC) that graphically represents the item's performance. Typically, an ICC is roughly S-shaped and different parts of the curve reveal information about each of the three parameters of interest.

Figure 7.3 shows what an ICC looks like. The X axis represents the strength of the characteristic or attribute being measured (e.g., knowledge, strength, accuracy, depression, social desirability, or virtually any other measurable phenomenon). The Y axis represents the probability of "passing" the item in question, based on the proportions of failing and passing scores observed. Seeing how an ICC can be used to assess item quality is actually easier if we look at a diagram representing two items that we can compare.

Figure 7.4 illustrates item difficulty by showing two curves. Note that the points at which the curves attain a 50% probability of their respective items being passed are different. For the lighter curve, that point is farther to the right. That is, the amount of the attribute must be higher for an individual to have a 50% chance of passing the item represented by the lighter line than the item represented by the darker line. Using that criterion, the item represented by the lighter line is thus more difficult. Difficulty in this case is not a subjective judgment but a factual description of the point on the X axis corresponding to the curve's crossing of the .50 probability value on the Y axis.

Figure 7.5 illustrates how we assess discrimination using the same two ICCs. The item corresponding to the darker curve has a steeper slope at the

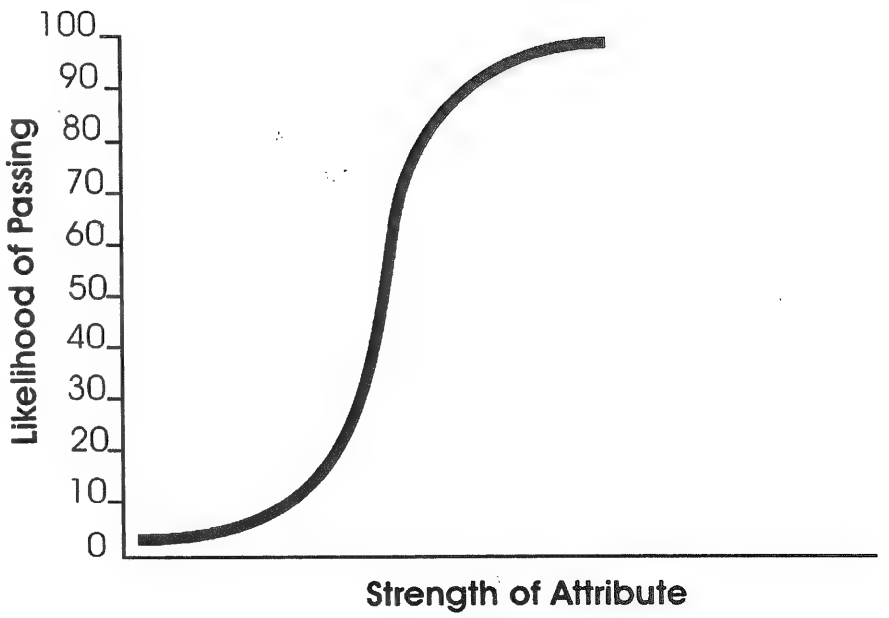

Figure 7.3 An example of an item characteristic curve (ICC)

50%-pass point than does the item represented by the lighter curve. A consequence of this is that it takes a smaller increase in the attribute to move from a clearly failing to a clearly passing score in the case of the item represented by the darker line. So, the steeper line of that item reveals that the region of the X axis that corresponds to an ambiguous score is smaller than the equivalent region for the other item. Thus, the dark-line item discriminates more effectively than the light-line item between those who fail and those who pass.

Finally, in Figure 7.6, we can look at the two items' propensities to yield passing grades even when the ability of the respondent (or whatever attribute of the respondent is being measured) is essentially zero. As you might have guessed, this is determined by the point at which the ICC intersects with the Y axis. For the dark-line item, the intercept value is close to zero. Thus, the probability of a person passing the item if he or she completely lacks the attribute in question is quite small. For the light-line item, there is a substantial probability (about 15%) that someone with no ability will pass the item and thus be indistinguishable, based on the item in question, from someone with very high ability. The corresponding diagram points out the differences in the

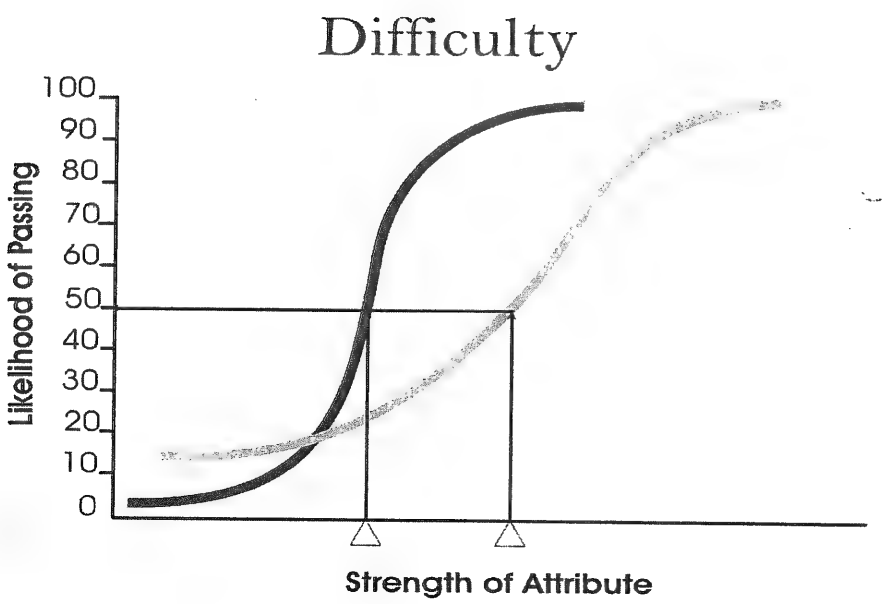

Figure 7.4 Examples of ICCs for two items that differ in difficulty

Y intercepts of the two items—the basis for concluding that, once again, the dark-line item is the better performer.

In theory, one can use IRT to establish the parameters for each of many items. Then, depending on the details of the intended application, optimally performing items can be used to address the measurement problem at hand. For example, "easy" items could be assembled for use with respondents who possess relatively low levels of the ability in question and difficult items assembled for those possessing high levels. This is directly analogous to using the 10-pound bell apparatus at a fair or carnival intended for children and the 100-pound apparatus at a camp for adult athletes. Using the wrong item—like using the wrong bell apparatus—can lead to either frustration (if the task is too difficult) or a lack of motivation (if the task is too easy). Also, if the measure being compiled will be the basis of important decisions, then minimizing the range of ambiguity for each item and the likelihood of false positives are also very attractive possibilities.

IRT approaches have the distinct advantage of directing our attention to three (in the three-parameter version that currently is popular) important aspects of an item's performance. With methods rooted in classical measurement theory,

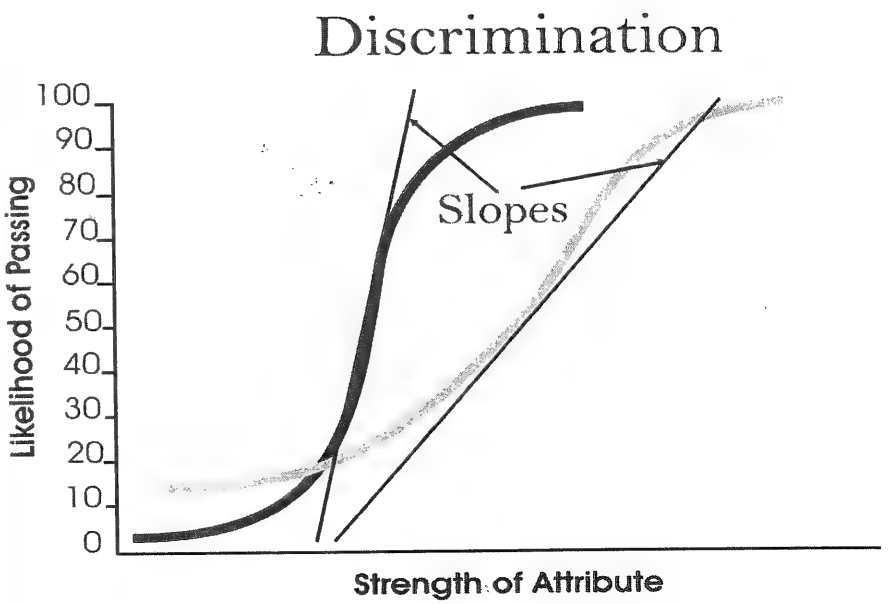

Figure 7.5 Examples of ICCs for two items that differ in their ability to discriminate the attribute they measure

we may know (e.g., from its performance in factor analysis or coefficient alpha computations) if an item is performing well or poorly, but we may not have as clear an understanding of the nature of any deficiencies it possesses. IRT, in contrast, may help us to assess an item's strengths and weaknesses more specifically.

COMPLEXITIES OF IRT

Although it is very attractive, IRT is not a quick solution to measurement problems. Like classical measurement theory, IRT does not determine the characteristics of items, it merely quantifies them. So, the technology per se allows a researcher to assess item performance but does not directly cause one to write better items or cause poorly constructed items suddenly to work well. Also, the assessment process can be daunting when the researcher is using methods based on IRT. Classical measurement trades precision for simplicity

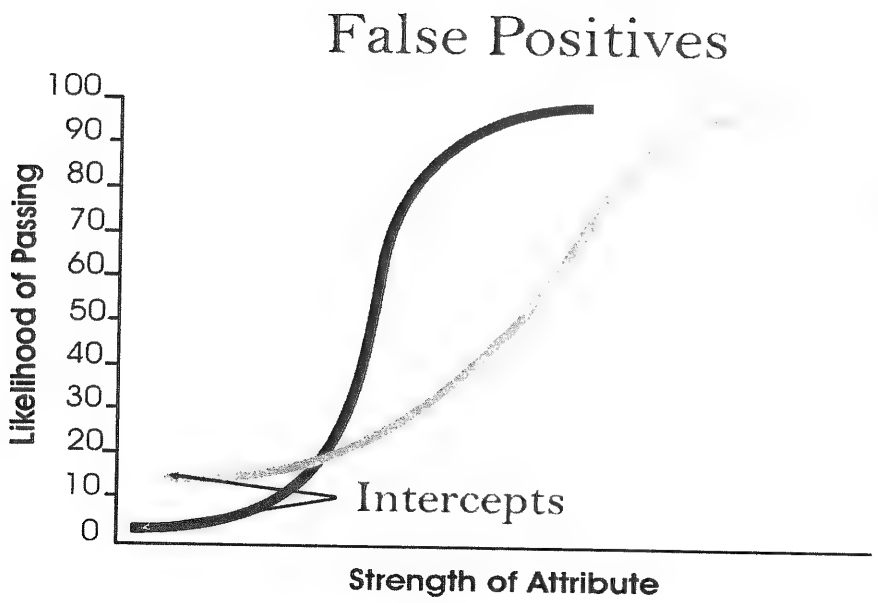

Figure 7.6 Examples of ICCs for two items that differ in their rates of false-positive responses

by adopting a less differentiated but more tractable conceptualization of error sources. IRT makes the opposite choice, gaining precision but sacrificing simplicity. Thus, IRT methods are demanding and have largely been confined to use by specialists. Up until the summer of 2002, most software for running IRT analyses was not Windows-based and had to be run, without a graphical interface, in the all-but-forgotten operating system DOS. The application of these methods, furthermore, requires a considerable degree of expert judgment. These methods are still in an active stage of development, with new issues popping up and new solutions being offered.

To have confidence that the characteristics of items are being assessed independently of the characteristics of the sample being studied, a primary objective of IRT, one must demonstrate that those characteristics are consistent across a variety of samples that differ in various respects, including ability level. It is important that the item characteristics not be associated with other attribute-independent sample characteristics, such as gender, age, or other variables that should be uncorrelated to the one being measured. Item scores should vary only when the attribute of interest varies, not because of

differences in any other variables. So, for example, if we assume that spelling ability is unrelated to gender, we would have to demonstrate that boys and girls with equal ability would have the same probability of passing an item. If this were not true, then gender or some other factor other than spelling ability would be influencing the item. Also, as with classical theory, items being examined in a set (i.e., making up an instrument for measuring the same variable) must share only a single underlying variable.

These requirements for people and items point out another potentially knotty issue: How does an instrument developer ascertain the true level of the attribute (usually called *theta*, signified as Θ) in a way that allows for the generation of the ICCs? Returning to our hammer-and-bell analogy, how do you define strength in order to determine how much strength it takes to ring the bell on a particular piece of equipment? In many cases, if the true level of the attribute is knowable in some manageable fashion, there would be little need to develop a new measure. In theory, given responses from a large number of people to a fixed set of items, a computer program should be able to sort out differences between the item and person characteristics. To return again to the analogy of the carnival devices (the bell-ringing apparatus and dunking machine), if enough people use, say, two of each device, it should be possible to determine which example of each type device is harder and, also, to judge the skills of individuals at these two tasks. In practice, there is often a back-and-forth, iterative process involving administering items to gauge the level of the attribute for particular respondents, then using that attribute estimate as a guide in determining the characteristics of other items. When the best items are identified on that basis, they can be used to obtain an improved estimate of individuals' levels of the attribute for the next round of item selection, and so on. Some methodologies rely on identifying *anchoring items* that demonstrably perform equivalently across groups and can serve as a basis for calibrating other items.

Given the nature of these processes, it is easy to see why IRT has been most enthusiastically embraced by commercial ability testing organizations such as those administering the Graduate Record Examinations. The constant administration and evaluation of items over time provides an excellent basis for finding items whose characteristics are stable across variation in a wide range of other respondent characteristics.

WHEN TO USE IRT

There are two situations in particular where the advantages of IRT approaches can be very important. Each is discussed below.

Hierarchical Items

The first of these involves items that are inherently hierarchical. Recall that, in classical theory, items are assumed to be roughly parallel indicators of the underlying latent variable. Each item is assumed to be approximately equivalent in its sensitivity to the phenomenon of interest. These assumptions fit well when assessing many personal characteristics, such as attitudes, beliefs, and mood states. For other variables, progression along a continuum is of primary interest and the items are structured accordingly. Measures of physical ability often conform to this pattern. For example, a measure of mobility using "yes" and "no" as response options might contain items assessing whether a respondent (a) can ambulate independently, (b) can ambulate only with an assistive device, or (c) is unable to walk at all. These items are not parallel. Each represents a different level of the attribute of interest. Accordingly, an IRT measurement model would probably fit better than a classical model in this situation.

Note that this is different from a situation in which someone is asked essentially parallel items but can use hierarchical responses to indicate endorsement. An example of the latter would be two items worded to capture roughly equivalent degrees of depression, accompanied by a 6-point agree-disagree response scale. In this case, you would expect the response option selected to be consistent across items for the same individual. With hierarchical items, such as those mentioned above describing mobility, you would not. In fact, responding "yes" to one (e.g., "is unable to walk at all") would be inconsistent with responding "yes" to another (e.g., "can ambulate independently"). Each item is, itself, tuned to a specific level of the attribute in a manner analogous to Thurstone or Guttman scaling, which I discussed in Chapter 5.

Another advantage of IRT with hierarchical items is that it may be possible to assemble item banks, with each item tuned to a particular range of ability. Items can then be chosen that apply to the situation at hand. This allows the test administrator to zero in on the appropriate attribute level, selecting items within the likely range of ability, and reducing the need for administering many out-of-range items. For example, if certain items have a consistently hierarchical relationship to one another, a test administrator could choose an item requiring a somewhat low level of the attribute being assessed and another requiring a somewhat high level. The specific choices could be based on an initial assessment of the respondent's ability level or on how the respondent answers an initial probe item. If a respondent passes the easy item and fails the difficult item, then only items with difficulties between those two need be considered for further administration. This is obviously more convenient than administering all items, from easiest to hardest. With IRT's ability

to calibrate items to ability levels, this form of administration is feasible (Jenkinson, Fitzpatrick, Garratt, Peto, & Stewart-Brown, 2001). Typically, this manner of item administration is computerized, and it is referred to as Computer Adaptive Testing (CAT; Van der Linden & Glas, 2000).

With the notable exception of cognitive abilities, psychological variables such as attitudes are not as frequently assessed by IRT methods as are physical variables such as health status. Some psychological variables may fit IRT models quite well, however. For example, self-efficacy is often assessed using hierarchical items, presenting progressively more challenging tasks or situations for the respondent to evaluate with respect to ease or confidence of execution (DeVellis & DeVellis, 2001). This would seem to be a situation in which IRT has a potential advantage. It is worth noting that measures of self-efficacy (and other variables) treated as if based on classical methods seem to work surprisingly well despite a hierarchical item structure and consequent lack of conformity to classical measurement assumptions.

Differential Item Functioning

A second situation in which IRT seems particularly advantageous is when it is important to distinguish differences in group membership and item characteristics. Such studies are concerned with *differential item functioning*, or DIF, that is, the tendency of an item to perform differently across groups that are actually equivalent with respect to the attribute being assessed. For example, if a test of depression administered to two different age groups showed a difference, it might be for either of two reasons or for both of them. That is, older and younger people might genuinely differ in depression or older and younger people may differ in their responses to the particular items administered, even when their levels of depression are similar. In order to compare two groups directly, one must assume that measures perform identically on both groups and that any differences observed are due only to the attribute of interest. This assumption is often reasonable but may require empirical verification when groups differ in some respects, such as culture or age, that might reasonably lead to between-group differences in item interpretation. IRT can be a powerful tool in such situations. Although it is difficult to gather the breadth of data necessary to use IRT methods across populations (such as ethnic groups), these methods provide far superior tools for assessing to what extent observed differences are due to item performance versus due to group differences. Classical methods may tell part of the story (e.g., by noting markedly different factor patterns across groups) but may not detect more subtle processes. Whether conclusions based on classical versus IRT models will diverge in a specific situation is an empirical question, but it is not

difficult to imagine this occurring. I suspect that, increasingly, it will be expected that IRT will be used for comparisons across socially relevant groupings of individuals.

It is worth noting that many real-world assessment situations entail both hierarchical items and potential opportunities for DIF. Educational assessment is one, but certainly not the only, such context in which IRT has been applied. Another particularly relevant example is the assessment of health outcomes. The endpoints of interest are often hierarchical. For example, how much pain, disability, or social disruption one experiences following a medical procedure vary along continua, and individual items may correspond to different points along those continua. Consequently, answering two items affirmatively will have different meanings, depending on where those items lie along the continuum in question. Some scoring system that recognizes and deals with the nonequivalence of individual items is required. IRT potentially can address this issue.

Furthermore, policy makers are keenly aware of health disparities across ethnic groups and would like the ability to quantify them accurately—to differentiate true group differences from DIF. Once again, IRT models seem particularly well-suited for such issues. Researchers using IRT-based measurement procedures have been quite active in this arena.

CONCLUSIONS

Measurement methods based on IRT have many attractive features. But developing good items is hard work, irrespective of the theoretical framework that guides the process. Writing items that consistently measure the attribute under study and are insensitive to other respondent characteristics is no mean feat. Whereas, in a measure based on classical theory, items can offset the imperfections of other items to a degree, the logic of IRT has each individual item standing and being judged on its own (although it is certainly possible to develop an instrument in which items presumably measure the same phenomenon but differ with respect to difficulty as we discussed it earlier). Because one *can* discover, by examining ICCs, for example, that an item performs well does not mean that one *will*. Having credible independent knowledge of the attributes being measured is a requirement of IRT that is difficult to satisfy strictly but that can be very adequately approximated with repeated testing of large and heterogeneous samples. When this is not an option, it may be exceedingly difficult to convince critics that the assumption has been adequately met.

My personal view is that where the assumptions of classical measurement theory are applicable, that is, where the items are intended as equivalent indicators of a common underlying variable, the tractability and performance of such measures make them attractive options. On the other hand, if the research question implies inherently hierarchical responses or concerns DIF, then the added complexity of IRT-based approaches may be the best choice. The mere use of those methods, however, is by no means a guarantee of the desired end product. The researcher must demonstrate that the assumptions of the chosen method have been met within acceptable limits and that the resultant measurement tool's reliability and validity are empirically verifiable.

Does IRT render classical methods obsolete? Many advocates of IRT recognize that both CMT and IRT have a role to play. For example, Embretson and Hershberger (1999), in the first of their recommendations for changes to current measurement approaches, state that "IRT and CTT approaches should be integrated in a comprehensive approach to measurement issues" (p. 252). Recently, as part of an expert panel convened to examine measurement models in the context of assessing health disparities, David Cella (2001) made the following observation:

> There is a tendency in the depths of the measurement field to argue various aspects of one approach over another, both across and within general traditions of measurement. Over time, I have been far more impressed with the commonality of approaches and, more important, the conclusions of approaches than with the differences. Classical test theory and item response theory may have sharp differences in the way they handle individual responses to questions, and the way a score is generated, but rarely do results generated under one tradition deviate markedly from those generated under the other. (p. 63)

While IRT will continue to increase in popularity, classical methods will coexist, much as regression analysis shares the stage with structural equation modeling (SEM) approaches. Although both IRT and SEM have carried things further than their antecedents, the earlier methods retain their utility.

8

Measurement in the Broader Research Context

The opening chapter of this volume set the stage for what was to follow by providing some examples of when and why measurement issues arise, discussing the role of theory in measurement, and emphasizing the false economy of skimping on measurement procedures. In essence, it sketched the larger context of research before the focus shifted to the specific issues covered in later chapters. This chapter returns to the big picture and looks briefly at the scale within the larger context of an investigation.

BEFORE SCALE DEVELOPMENT

Look for Existing Tools

Very early in this volume, I suggested that scale development often is the result of a lack of appropriate existing instruments. It is important and efficient to be quite certain that a suitable measurement alternative does not already exist. Elsewhere (DeVellis, 1996) I have suggested ways of searching for appropriate scales. Essentially, this process involves searching published and electronic compendia of measures to determine whether a suitable instrument already exists. Published series such as the *Mental Measurements Yearbook* (e.g., Kramer & Conoley, 1992) and *Tests in Print* (e.g., Murphy, Conoley, & Impara, 1994) contain primarily clinical measures, including tests of ability and personality. These are often the tools that applied psychologists will use for assessing clients. Tools intended primarily for research are less prominent but some are included in these series. Another category of resources is targeted compilations, such as *Measures of Personality and Social Psychological Attitudes* (Robinson, Shaver, & Wrightsman, 1991). Relevant journals are also an excellent place to find which measurement strategies have worked successfully for others interested in the same constructs.

With increasing frequency, compilations of information about measurement instruments are being placed on the World Wide Web. In fact, the Web may be where the most rapid expansion of measurement-related information is

appearing. Internet sites oriented around a specific research topic (e.g., veterans, elderly persons, minority issues) sometimes include bibliographies of measurement tools used in that type of research. Some sites are specifically intended to provide assistance in identifying appropriate measurement tools.

One especially useful Web-accessible resource is the Measurement Excellence Initiative (MEI), a program of the Department of Veterans' Affairs. Their site is both a repository for information related to measurement and a gateway to other print- and Web-based compendia of measurement tools and information. Although its primary focus is measurement related to health services research, the site also includes information about measurement theories and tools with broader applications. It is located at www.measurementexperts.org. One of the many links from the MEI site is the Health and Psychological Instruments (HaPI) database (Behavioral Measurement Database Services, 2000), which is also available online through many university libraries. It contains information on a large (literally thousands) and growing collection of instruments used in research. Typically, the article in which an instrument was first published is abstracted. In addition, information from subsequent applications of the instrument that contain relevant psychometric information may also be included. The depth of information depends largely on what was made available to HaPI. Consequently, some measures are not described in depth while others are. Despite certain limitations, it can be a valuable resource for identifying potentially relevant instruments.

As with any Web-based information, the consumer needs to consider the origin of the information and its trustworthiness. Sites sponsored by universities, government agencies (such as the MEI site), and other established institutions or organizations (such as HaPI) usually have accurate and credible information. The MEI site reviews all others to which it provides links and includes only those that it considers credible and responsible. In general, however, caution is warranted. Just as there are an abundance of "junk science" books, there are also plenty of internet sites that may adopt a tone and appearance of scientific legitimacy that are not commensurate with their content. The skills you have gained from this book should help you to evaluate measurement information sources more critically in any format and to determine whether the measures described have demonstrated adequate reliability and validity.

View the Construct in the Context of the Population of Interest

We have already discussed the importance of theoretical clarity. It is often important to assess whether the theoretical constructs we as researchers

identify correspond to the actual perceptions and experiences of the people we plan to study. Focus groups (see, e.g., Krueger & Casey, 2000) can be a means of determining whether ideas that underlie constructs of interest make sense to respondents. As an example, consider attributions—the explanations or interpretations people ascribe to various outcomes. Often these are assessed along dimensions such as "under my control" versus "not under my control," "specific to this situation" versus "applicable to most situations," and "a characteristic of me" versus "a characteristic of the environment or situation." Research into attribution processes has been highly productive. Most people can analyze outcomes, such as receiving a job offer following an interview, along such dimensions. In some situations, however, this may not be the case. For example, asking foreign, rural-dwelling, poorly educated elders, unaccustomed to thinking in such terms, to evaluate illness outcomes or purchasing decisions along these three dimensions may not work well. Experience suggests that they may simply not understand the task because it is so foreign to the way they think about things. A focus group asking potential research participants to discuss relevant concepts might make such a problem evident and preclude a measurement strategy that would be doomed to failure.

Focus groups can also reveal the natural, everyday language that people use to talk about a concept. A young mother may not use the same terms as a marketing expert to describe reactions to a product. The former may use "pretending" to describe her child playing without using a particular toy, whereas a marketing researcher might describe such play as "product nondirected." Structuring items around her language usage (e.g., "How much playtime does your child spend just pretending, without using any toys?") rather than the expert's (e.g., "How much time does your child spend in product nondirected play?") is more likely to yield a tool suitable for measuring her perceptions of how her child interacts with various products.

A note of caution: Some researchers advocate having the target population give final approval of questionnaires. This is admirable and is likely to give the participants a greater sense of active participation in the research process. However, it is unfair to expect nonexperts to understand technical issues that apply to item construction, as discussed in Chapter 5. For example, a nonexpert might prefer an item that is worded in a pleasant, moderate way, while an experienced scale developer might recognize that the preferred wording would generate little variation in responses, thus rendering the item useless. My personal recommendation is to help participants feel actively involved in a variety of ways, if that is appropriate, but to reserve the right of final approval on item wording. We do not honor our research participants if we inadvertently create a situation in which their views, feelings, or opinions cannot be accurately gauged. We simply waste their time.

There are other methods that also can be used to determine whether participants understand questions in the way intended. For example, simply asking people what they understand a question to mean, or to think aloud while forming an answer, may be highly effective during instrument pilot testing. The more general point is to understand who the respondents will be and to determine which ways of expressing concepts will be most clear to them.

Decide on the Mode of Scale Administration

Researchers can collect data in a variety of ways (e.g., Dillman, 2000), and they may choose to match modes of administration to the preferences of respondents. Accordingly, a team of investigators might consider using an interview rather than a printed questionnaire. It is important to recognize that a scale intended to be completed in print form may have substantially different properties when the items and responses are presented orally. For example, parents might be more reluctant to acknowledge high aspirations for their children if they had to report them aloud to an interviewer rather than marking a response option. (Investigators contemplating data collection by modes other than self-administered questionnaires should consult volumes in the Applied Social Research Methods series by Lavrakas, 1993, and Fowler and Mangione, 1989.) Generally, it is wise to restrict the mode of administering a new scale to the method used during scale development. A G-study (see Chapter 3) may be used to determine the scale's generalizability across administration modes.

Consider the Scale in the Context of Other Measures or Procedures

What questions or research procedures will precede the scale itself? How will these questions affect responses to the scale? Nunnally (1978, pp. 627-677) refers to contextual factors such as response styles, fatigue, and motivation as *contingent variables*. He also points out that they can adversely affect research in three ways: (a) by reducing the reliability of scales; (b) by constituting reliable sources of variation other than the construct of interest, thus lowering validity; and (c) by altering the relationships among variables, making them appear, for example, to be more highly correlated than they actually are. As an example of how contingent variables might operate, consider *mood induction* and *cognitive sets* as they might apply to the marketing research example. The former might be an issue if, for example, the market researchers decided to include a depression or self-esteem scale in the same questionnaire as their aspirations scale. Scales tapping these (and other) constructs often contain items that express negative views of one's self. The

Rosenberg Self-Esteem scale (Rosenberg, 1965), for example, contains such items as "I feel I do not have much to be proud of" (as well as items expressing a positive self-appraisal). A researcher who was not sensitive to the potential effects of mood induction might select a series of exclusively self-critical items to accompany a newly developed scale. Reading statements that consistently express negative assessments of one's self may induce a dysphoric state that, in turn, can cause whatever follows to be perceived differently than it would otherwise have been perceived (Kihlstrom, Eich, Sandbrand, & Tobias, 2000; Rholes, Riskind, & Lane, 1987). This might have each of the three adverse effects noted by Nunnally. That is, in the presence of affectively negative items, aspiration items might take on a somewhat different shade of meaning, thus lowering the proportion of variance in those items that is attributable to the intended latent variable. Or, in an extreme instance, some items from the aspiration scale might come to be influenced primarily by the induced mood state, rendering the scale multifactorial and lowering its validity as a measure of parental aspiration. Finally, to the extent that respondents' mood affected their responses to the aspiration items, scores for this scale might correlate artificially highly with other mood-related measures.

Cognitive sets are a more general example of the same phenomenon; that is, some frame of reference other than mood might be induced by focusing respondents' attention on a specific topic. For example, immediately preceding the aspiration scale with items concerning the respondents' income, the value of their home, and the amount they spend annually on various categories of consumer goods might create a mental set that temporarily altered their aspirations for their children. As a result, the responses to the scale might reflect an unintended transient state. As with mood, this cognitive set might adversely affect the reliability and/or validity of the scale by contaminating the extent to which it unambiguously reflects parental aspirations.

AFTER SCALE ADMINISTRATION

A quite different set of issues emerges after the scale has been used to address a substantive research question. The primary concerns at this point are the analysis and interpretation of the data generated by the instrument.

Analytic Issues

One issue in data analysis is the appropriateness of various techniques for variables with different scaling properties. The theoretical perspective and methods advocated most strongly in this book should result in scales that are

amenable to a wide variety of data analytic methods. Although, strictly speaking, items using Likert or semantic differential response formats may be ordinal, a wealth of accumulated experience supports applying interval-based analytic methods to the scales they yield. However, the question of which methods are best suited to which types of data has been, and certainly will continue to be, hotly debated in the social sciences. Determining how different response options affect estimates of underlying variables is an active research area in its own right. Also, different audiences will have different expectations for how measures are treated. Whereas psychologists, for example, may be fairly sanguine with treating Likert scales as producing interval-level data, epidemiologists may not be. Perhaps the most practical approach is to monitor (and conform to) the prevailing sentiment with respect to this issue in one's area of interest.

Interpretation Issues

Assuming that the researcher has arrived at a suitable analytic plan for the data generated by a newly developed scale, the question of how to interpret the data remains. One point to keep in mind at this juncture is that the validity of a scale is not firmly established during scale development. Validation is a cumulative, ongoing process. Moreover, validity is really a characteristic of how a scale is used, not of the scale itself. A depression scale, for example, may be valid for assessing depression but not for assessing general negative affect.

Also, it is important to *think* about one's findings. Especially if the results appear strongly counterintuitive or countertheoretical, the researcher must consider the possibility that the scale is invalid in the context of that particular study (if not more broadly). It may be that the extent to which the validity of the scale generalizes across populations, settings, specific details of administration, or an assortment of other dimensions, is limited. For example, the hypothetical parental aspiration measure might have been developed with a relatively affluent population in mind, and its validity for individuals whose resources are more limited may be unacceptable. Any conclusions based on a scale that has had limited use should consider the following: (a) how its present application differs from the context of its original validation, (b) the likelihood that those differences might limit the validity of the scale, and (c) the implications of those limitations for the present research.

Generalizability

The previous paragraph cautioned against generalizing across populations, settings, and other aspects of research. This issue warrants further emphasis. Reaching conclusions about group differences potentially confounds

differences in the phenomenon being measured and differences in the performance of the instrument. If we can assume the latter to be trivial, then we can ascribe observed differences to group membership. In many situations (e.g., comparing off-task time in two groups of children randomly selected and allocated to groups), this will be the case. In others (e.g., comparisons across culturally distinct groups), we cannot assume identical instrument performance. DIF, discussed in Chapter 7, is an active area of psychometric research. Although most researchers will not make it a focus of their own efforts, they should recognize the possibility of its presence and the limitations it may impose on their conclusions.

FINAL THOUGHTS

Measurement is a vital aspect of social and behavioral research. No matter how well designed and executed other aspects of a research endeavor may be, measurement can make or break a study. We assume that the variables of interest to us correspond to the assessment procedures we use. Often, the relationship of primary interest is between two or more unobservable variables, such as desire for a certain outcome and failure to consider alternative outcomes. We cannot directly measure desire or consideration, so we construct measures that we hope will capture them. These measures are, in a sense, quantitative metaphors for the underlying concepts. Only to the extent that those metaphors are apt (i.e., the instruments are valid) will the relationships we observe between measures reflect the relationship we wish to assess between unobservable constructs. Exquisite sampling, superb research design, and fastidious implementation of procedures will not change this fact. A researcher who does not understand the relationship between measures and the variables they represent, in a very literal sense, does not know what he or she is talking about. Viewed in this light, the efforts entailed in careful measurement are amply rewarded by their benefits.

References

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting behavior.* Englewood Cliffs, NJ: Prentice Hall.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole.

Anastasi, A. (1968). *Psychological testing* (3rd ed.). New York: Macmillan.

Asher, H. B. (1983). *Causal modeling* (2nd ed.). Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07-003. Beverly Hills, CA: Sage.

Barnette, W. L. (1976). *Readings in psychological tests and measurements* (3rd ed.). Baltimore, MD: Williams & Wilkins.

Behavioral Measurement Database Services. (2000). *Ovid Technologies field guide: Health and Psychological Instruments (HAPI).* Retrieved May 1, 2002, from http://www.ovid.com/documentation/user/field_guide/disp_fldguide.cfm?db=hapidb.htm

Blalock, S. J., DeVellis, R. F., Brown, G. K., & Wallston, K. A. (1989). Validity of the Center for Epidemiological Studies Depression scale in arthritis populations. *Arthritis and Rheumatism, 32,* 991-997.

Bohrnstedt, O. W. (1969). A quick method for determining the reliability and validity of multiple-item scales. *American Sociological Review, 34,* 542-548.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: John Wiley.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Carmines, E. G., & McIver, J. P. (1981). Analyzing models with unobserved variables: Analysis of covariance structures. In G. W. Bohrnstedt & B. F. Borgatta (Eds.), *Social measurement: Current issues* (pp. 65-115). Beverly Hills, CA: Sage.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1,* 245-276.

Cella, D. (2001, May). Commentary provided at the Resource Centers on Minority Aging Research conference on *Measurement Issues in Health Disparities Research in the U.S.*, San Francisco.

Cohen, P., Cohen, J., Teresi, J., Marchi, M., & Velez, C. N. (1990). Problems in the measurement of latent variables in structural equation causal models. *Applied Psychological Measurement, 14,* 183-196.

Comrey, A. L. (1973). *A first course in factor analysis.* New York: Academic Press.

Comrey, A. L. (1988). Factor analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology, 56,* 754-761.

Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire.* Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07-063. Beverly Hills, CA: Sage.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart & Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *Dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: John Wiley.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Cureton, E. E. (1983). *Factor analysis: An applied approach.* Hillsdale, NJ: Lawrence Erlbaum.

Currey, S. S., Callahan, L. F., & DeVellis, R. F. (2002). *Five-item Rheumatology Attitudes Index (RAI): Disadvantages of a single positively worded item.* Unpublished paper, Thurston Arthritis Research Center, University of North Carolina at Chapel Hill.

Czaja, R., & Blair, J. (1996). *Designing surveys: A guide to decisions and procedures.* Thousand Oaks, CA: Pine Forge.

Dale, F., & Chall, J. E. (1948). A formula for predicting readability: Instructions. *Education Research Bulletin, 27,* 37-54.

DeVellis, B. M., & DeVellis, R. F. (2001). Self-efficacy and health. In A. Baum & T. Revenson (Eds.), *Handbook of health psychology.* Mahwah, NJ: Lawrence Erlbaum.

DeVellis, R. F. (1996). A consumer's guide to finding, evaluating, and reporting on measurement instruments. *Arthritis Care and Research, 9,* 239-245.

DeVellis, R. F., Blalock, S. J., Holt, K. D., Renner, B. R., Blanchard, L. W., & Klotz, M. L. (1991). Arthritis patients' reactions to unavoidable social comparisons. *Personality and Social Psychology Bulletin, 17,* 392-399.

DeVellis, R. F., & Callahan, L. F. (1993). A brief measure of helplessness: The helplessness subscale of the Rheumatology Attitudes Index. *Journal of Rheumatology, 20,* 866-869.

DeVellis, R. F., DeVellis, B. M., Blanchard, L. W., Klotz, M. L., Luchok, K., & Voyce, C. (1993). Development and validation of the Parent Health Locus of Control (PHLOC) scales. *Health Education Quarterly, 20,* 211-225.

DeVellis, R. F., DeVellis, B. M., Revicki, D. A., Lurie, S. J., Runyan, D. K., & Bristol, M. M. (1985). Development and validation of the child improvement locus of control (CILC) scales. *Journal of Social and Clinical Psychology, 3,* 307-324.

DeVellis, R. F., Holt, K., Renner, B. R., Blalock, S. J., Blanchard, L. W., Cook, H. L., Klotz, M. L., Mikow, V., & Harring, K. (1990). The relationship of social comparison to rheumatoid arthritis symptoms and affect. *Basic and Applied Social Psychology, 11,* 1-18.

Dillman, D. A. (2000). *Mail and internet surveys: The tailored design method* (2nd ed.). New York: John Wiley.

Duncan, O. D. (1984). *Notes on social measurement: Historical and critical.* New York: Russell Sage.

Embretson, S. E., & Hershberger, S. L. (1999). Summary and future of psychometric models in testing. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement* (pp. 243-254). Mahwah, NJ: Lawrence Erlbaum.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7,* 117- 140.

Fink, A. (1995). *The survey kit.* Thousand Oaks, CA: Sage.

Fowler, F. J. (1993). *Survey research methods.* Thousand Oaks, CA: Sage.

Fowler, F. J. (1995). *Improving survey questions: Design and evaluation.* Thousand Oaks, CA: Sage.

Fowler, F. J., & Mangione, T. W. (1989). *Standardized survey interviewing.* Newbury Park, CA: Sage.

Fry, E. (1977). Fry's readability graph: Clarifications, validity, and extension to level 17. *Journal of Reading, 21,* 249.

Ghiselli, B. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* San Francisco: Freeman.

Gorsuch, R. L. (1983). *Factor analysis.* Hillsdale, NJ: Lawrence Erlbaum.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Harman, H. H. (1976). *Modern factor analysis.* Chicago: University of Chicago Press.

Hathaway, S. R., & McKinley, J. C. (1967). *Minnesota Multiphasic Personality Inventory: Manual for administration and scoring.* New York: Psychological Corporation.

Hathaway, S. R., & Meehl, P. E. (1951). *An atlas for the clinical use of the MMPI.* Minneapolis: University of Minnesota Press.

Idler, E. L., & Benyamini, Y. (1997). Self-rated health and mortality: A review of twenty-seven community studies. *Journal of Health and Social Behavior, 38,* 21-37.

Jenkinson, C., Fitzpatrick, R., Garratt, A., Peto, V., & Stewart-Brown, S. (2001). Can item response theory reduce patient burden when measuring health status in neurological disorders? Results from Rasch analysis of the SF-36 physical functioning scale (PF-10). *Journal of Neurology, Neurosurgery, and Psychiatry, 71,* 220-224.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 109-134.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement 20,* 141-151.

Keefe, F. J. (2000). Self-report of pain: Issues and opportunities. In A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report: Implications for research and practice* (pp. 317-337). Mahwah, NJ: Lawrence Erlbaum.

Kelly, J. R., & McGrath, J. B. (1988). *On time and method.* Newbury Park, CA: Sage.

Kihlstrom, J. F., Eich, E., Sandbrand, D., & Tobias, B. A. (2000). Emotion and memory: Implications for self-report. In A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report: Implications for research and practice* (pp. 81-103). Mahwah, NJ: Lawrence Erlbaum.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). San Francisco: Brooks/Cole.
Kramer, J. J., & Conoley, J. C. (1992). *The eleventh mental measurement yearbook.* Lincoln, NE: Boros Institute of Mental Measurements.
Krueger, R. A., & Casey, M. A. (2000). *Focus groups: A practical guide for applied research.* Thousand Oaks, CA: Sage.
Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision.* Sage Applied Social Research Methods Series, Vol. 7. Thousand Oaks, CA: Sage.
Levenson, H. (1973). Multidimensional locus of control in psychiatric patients. *Journal of Consulting and Clinical Psychology, 41,* 397-404.
Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research.* Newbury Park, CA: Sage.
Loehlin, J. C. (1998). *Latent variable models: An introduction to factor, path, and structural analysis.* Mahwah, NJ: Lawrence Erlbaum.
Long, J. S. (1983). *Confirmatory factor analysis.* Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07-033. Beverly Hills, CA: Sage.
MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4,* 84-99.
Mayer, J. M. (1978). Assessment of depression. In M. P. McReynolds (Ed.), *Advances in psychological assessment* (Vol. 4, pp. 358-425). San Francisco: Jossey-Bass.
McDonald, R. P. (1984). *Factor analysis and related methods.* Hillsdale, NJ: Lawrence Erlbaum.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.
Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin, 86,* 376-390.
Murphy, L. L., Conoley, J. C., & Impara, J. C. (1994). *Tests in print IV.* Lincoln, NE: Boros Institute of Mental Measurements.
Myers, J. L. (1979). *Fundamentals of experimental design* (3rd ed.). Boston: Allyn & Bacon.
Namboodiri, K. (1984). *Matrix algebra: An introduction.* Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07-028. Beverly Hills, CA: Sage.
Narens, L., & Luce, R. D. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin, 99,* 166-180.
Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
Osgood, C. E., & Tannenbaum, P. H. (1955). The principle of congruence in the prediction of attitude change. *Psychological Bulletin, 62, 42-55.*
Radloff, L. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1,* 385-401.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: MESA Press.
Rholes, W. S., Riskind, J. H., & Lane, J. W. (1987). Emotional states and memory biases: Effects of cognitive priming and mood. *Journal of Personality and Social Psychology, 52,* 91-99.
Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (1991). *Measures of personality and social psychological attitudes.* San Diego, CA: Academic Press.
Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton, NJ: Princeton University Press.
Rotter, J. B. (1966). Generalized expectancies for internal vs. external control of reinforcement. *Psychological Monographs, 80* (1, Whole No. 609).
Saucier, G., & Goldberg, L. R. (1996). The language of personality: Lexical perspectives on the five-factor model. In J. S. Wiggins (Ed.), *The five factor model of personality* (pp. 21-50). New York: Guilford.
Smith, P. H., Earp, J. A., & DeVellis, R. F. (1995). Measuring battering: Development of the Women's Experiences with Battering (WEB) scale. *Women's Health: Research on Gender, Behavior, and Policy, 1,* 273-288.
Spielberger, C. D., Gorsuch, R. L., & Lushenc, R. E. (1970). *State-trait anxiety inventory (STAI) test manual for form X.* Palo Alto, CA: Consulting Psychologists Press.
Strahan, R., & Gerbasi, K. (1972). Short, homogenous version of the Marlowe-Crowne Social Desirability Scale. *Journal of Clinical Psychology, 28,* 191-193.
Tinsley, H. E. A., & Tinsley, D. J. (1987). Uses of factor analysis in counseling psychology research. *Journal of Counseling Psychology, 34,* 414-424.
Van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice.* St. Paul, MN: Assessment Systems.
Wallston, K. A., Stein, M. J., & Smith, C. A. (1994). Form C of the MHLC Scales: A condition-specific measure of locus of control. *Journal of Personality Assessment, 63,* 534-553.
Wallston, K. A., Wallston, B. S., & DeVellis, R. (1978). Development and validation of the multidimensional health locus of control (MHLC) scales. *Health Education Monographs, 6,* 161-170.
Weisberg, H., Krosnick, J. A., & Bowen, B. D. (1996). *An introduction to survey research, polling, and data analysis.* Thousand Oaks, CA: Sage.
Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.) *The new rules of measurement* (pp. 65-104). Mahwah, NJ: Lawrence Erlbaum.
Zorzi, M., Priftis, K., & Umilitá, C. (2002). Brain damage: Neglect disrupts the mental number line. *Nature, 417* (09 May), 138-139.
Zuckerman, M. (1983). The distinction between trait and state scales is not arbitrary: Comment on Allen and Potkay's "On the arbitrary distinction between traits and states." *Journal of Personality and Social Psychology, 44,* 1083-1086.
Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99,* 432-442.

Index

About the Author

Robert F. DeVellis is Research Professor in the Department of Health Behavior and Health Education (School of Public Health), and the Psychology Department (College of Arts and Sciences) at the University of North Carolina at Chapel Hill. In addition, he is a Core Faculty Member for UNC's Robert Wood Johnson Clinical Scholars Program (School of Medicine). Dr. DeVellis is also Director of the Measurement and Methods Core of the UNC Center on Minority Aging and Associate Director of the UNC Arthritis Multidisciplinary Clinical Research Center, where he also is a member of that center's Methodology Core. He has served on the Board of Directors for the American Psychological Association's Division of Health Psychology (38), on the Arthritis Foundation's Clinical/Outcomes/Therapeutics Research Study Section, and on the Advisory Board of the Veterans Affairs Measurement Excellence Initiative. He has served on the editorial boards of *Arthritis Care and Research* and *Health Education Research* and as Guest Editor, Guest Associate Editor, or reviewer for more than two dozen other journals. His current research interests include examining interpersonal factors that facilitate adaptation to chronic illness and measuring social and behavioral variables related to health and illness. He has served as Principal Investigator or Co-Investigator since the early 1980s on a series of research projects funded by the federal government and private foundations.